Local fit in multilevel latent class and hidden Markov models

Nagelkerke, Erwin

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Local Fit

In Multilevel Latent Class & Hidden Markov Models

Erwin Nagelkerke

Local Fit in Multilevel Latent Class and Hidden Markov Models

ERWIN NAGELKERKE

Tilburg University
School of Social and Behavioral Sciences
Methodology & Statistics

This research is funded by The Netherlands Organization for Scientific Research (NWO) [Vici grant number 453-10-002].

Printing was financially supported by Tilburg University.

Printed by:	ProefschriftMaken, Vianen, The Netherlands
Cover inspired by:	Lianne Ippel

Local Fit in Multilevel Latent Class and Hidden Markov Models

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 16 februari 2018 om 14.00 uur

door

Erwin Nagelkerke,

geboren op 18 november 1986
te Kapelle.

Promotor:	prof. dr. J. K. Vermunt
Copromotor:	dr. D. L. Oberski
Overige leden van de Promotiecommissie:	prof. dr. C. A. W. Glas
	dr. E. L. Hamaker
	dr. L. C. J. M. Halman
	dr. Z. Bakk

Contents

Chapter 1

Introduction

In many scientific fields certain concepts or characteristics are used that are not directly observable. Examples of these are plentiful, in science as well as daily life, since many descriptions of people, objects, organizations, and events include difficult to assess or broad concepts. A prime example of this is intelligence. When describing someone as smart whilst telling a story at a party, this description is often based on different observations of that person where he or she may have answered questions correctly during trivia, got high grades in college, opted a creative solution to a problem, or talked about extensive responsibilities at their job. Fortunately, it is unlikely that people listening to the story will demand an explanation of how this characteristic was measured and which observations played a role in coming to the conclusion that smart is indeed a good description that has some truth to it.

In science such objectivity generally is required, and as a result methods have been developed that measure such an unobservable (latent) phenomenon by combining multiple (manifest) measurements that could be made and are indicative of the unobserved characteristic. A very well-known method is that of factor analysis, which uses a number of observed variables to construct a score on the latent variable. For example, by combining many test items that measure language proficiency and mathematical skills an IQ score can be constructed. Item response theory is similar, but also attempts to distinguish between the difficulty of the test items and the ability of the respondent. However, sometimes not one continuous value for a certain characteristic, but a categorization is needed. For example in cases where a typology such as personality type or social-economic class is measured, or in the case of diagnoses where respondents need to be classified according to having a certain illness or not. In these cases where the latent variable is categorical, and often many of the manifest indicator variables such as the presence or absence of symptoms as well, latent class analysis is a very general and broadly applicable method.

Whether or not these statistical models provide a good description of the latent concept depends on a range of issues: finding a representative set of indicators, making sure that the observations are a representative sample of the population of interest, applying the right statistical models, and applying them correctly. Here the focus is on the latter two, namely the suitability of latent class and latent Markov models to describe the data at hand.

1.1 The Latent Class Model and Extensions

Latent class (LC) analysis was originally developed and demonstrated by Lazarsfeld in 1950 (Lazarsfeld, 1950, 1959) as a probabilistic approach to model psychometric, binary data. Work that he formalized and extended in 1968 (Lazarsfeld & Henry, 1968), which fifty years later is still a comprehensive and useful introduction to the LC model. Half a decade later, Goodman (1974) extended the model to be applicable to nominal items and solved many problems associated with its estimation, introducing the basic LC model as it is used today.

In the social sciences LC analysis is generally used to classify respondents into unobserved, unknown groups based on their responses to usually categorical, observed variables. That is, based on their pattern of responses, respondents have a certain probability to belong to a certain category on a latent variable. Some examples of this are distinguishing behavior patterns, such as combinations and severity of adolescent substance use (Gilreath et al., 2014), creating a typology based on personal characteristics, such as categorizing households into social economic classes based on income and social status (Savage et al., 2013), or classifying patients based on illness manifestations, such as the severity and comorbidity of depressive symptoms (Ferdinand, De Nijs, Van Lier, & Verhulst, 2005).

This model, like many other statistical models, assumes the observations in the data to be independent, which is problematic in the case of complex sampling designs. For example, respondents may be observed in naturally occurring, manifest groups or respondents may be observed at multiple different times. In those cases the assumed independence of observations may not hold, as people from the same group, or observations of the same person, tend to be systematically more similar to each other than those from different groups or persons (Hox, 2010, pp. 4-5). Ignoring this structure of the data will lead to biased results in the LC model as well (Kaplan & Keller, 2011; Park & Yu, 2016).

In order to take into account this similarity of members from the same group Vermunt proposed the multilevel LC model (Vermunt, 2003), which introduces random effects that allow observations from different groups to have different latent classes and different probabilities of belonging to those classes. An earlier approach to this was the multiple-group LC model (Clogg & Goodman, 1985), which estimates the LC model separately for each observed group. Because this results in large numbers of parameters that become unfeasible to interpret and compare it quickly loses its value when many groups are observed. However, therein also lies the key idea of the multilevel extension, because when many groups are observed it becomes possible to estimate the parameter distribution of the group-specific coefficients. That is, add a random-effects mixture component to the original model with a latent variable on the group level, in addition to the latent variable on the lower level.

The substantive benefit of this approach is that, in addition to the regular classification of respondents, it simultaneously allows a classification of the observed

groups (Vermunt, 2008). Examples of these are classifying students and the schools they attend in terms of (un-)healthy behavior (Allison, Adlaf, Irving, Schoueri - Mychasiw, & Rhem, 2016), classifying residents and countries according to preferred ways to purchase goods and services (Dal Bianco, Paccagnella, & Varriale, 2016), and creating a typology of the relation between team supervisors and their team members (Zinn, 2015).

Taking the dependence between repeated measurements of the same person into account, rather than that of members to the same group, seems to be not that different in terms of the structure of the data. Yet, the latent (or hidden) Markov (LM) model that is often used as a solution has developed more or less in parallel with the LC framework and the two have only later been reconciled. This is presumably mostly due to the goals of the initial developments. Wiggins (1955) originally developed the LM model to take into account measurement error for a single item that is measured multiple times for the same person, which he illustrated far more elaborately some years later (Wiggins, 1973).

The extensions to this original model took somewhat of a reverse course when compared to the multilevel LC model, since a way to take into account the nested structure of the data was present from the beginning in the form of a (first order) Markov chain, and it is the substantive goal of clustering that was added through several extensions. Most notably, after Baum, Petrie, Soules, & Weiss (1970) made it possible to efficiently estimate the model, and many contributions in the field of item response theory (e.g. Rasch, 1960; Birnbaum, 1968), the ideas behind LC and LM modeling were combined by allowing multiple indicators to measure the latent variables, (a.o. Poulsen, 1982; Van de Pol & Langeheine, 1990; Langeheine & Van de Pol, 1990), which implies that now a classification can be obtained similar to that of LC modeling, and can be combined with the Markov chain to allow respondents to switch between classes over time. In technical terms a finite mixture of Markov chains could now be estimated. Further extensions quickly followed, such as allowing covariates to be included (Vermunt, Langeheine, & Böckenholt, 1999; Bartolucci, Pennoni, & Francis, 2007).

Substantively this model allows, in a relatively parsimonious way, the respondents to be classified into states and the transition in and out of those states between different measurement occasions to be modeled. That is, the model also describes the structural change over time in the latent category that respondents belong to. The number of applications and further developments of the LM model are plentiful, and due to its popularity in the emerging data science and data mining fields are quickly growing. Some examples include detecting different types of students and their path to graduation or non-completion of college (Witteveen & Attewell, 2017), patterns and developments in criminal behavior (Bartolucci et al., 2007), and an enormous body of work in the area of speech recognition.

1.2 Goodness-of-Fit in the Multilevel Latent Class Model

Taking into account the additional dependence that results from complex sampling designs does mean that the multilevel LC and LM models become quite complex models with increased numbers of parameters and assumptions. Adherence to these assumptions and the correct estimation of the parameters is central to how well the model is able to summarize and capture the most important aspects of the observed data. In other words, how well the model fits to the data.

Arguably one of the most influential works in this area is by Pearson (1900), in which the chi-squared residual is described along with its asymptotic properties. Assuming that the sample is a correct representation of the population, the idea is that it should not be too unlikely that the differences between the predictions that follow from a system of equations, a statistical model, and the actually observed data are random errors. That is, when the predicted value is subtracted from the observed value this is a quantification of error, and these errors should be attributable to random chance instead of the model being outright wrong. This idea is so fundamental to statistics that essentially any model fit test does something similar.

The problem with inspecting the full system of equations at once, which is done with the chi-squared test (χ^2), the likelihood-ratio test (G^2), and adaptations of these like the Aikaike Information Criterion and Bayesian Information Criterion (see e.g. Burnham & Anderson, 2004), is that they can only state that something is wrong with the model. The latter AIC and BIC can even only be used to compare two models and indicate which is better. The problem with these global fit statistics is that modern complex models need to adhere to a range of assumptions to work, and have several substantive goals that are interwoven. It would then be useful to know whether the model adheres to each of the assumptions individually and to what extent it achieves its goals. This is especially true when considering that a model can overall fit relatively well to the data, but simultaneously have extensive misfit in one particular area. When that area is of interest to the research question, the conclusions will be biased, or wrong, without any way to detect this.

For the original LC model such a local fit statistic exists, that uses the idea of Pearson, and tests one central assumption that the model makes, namely that of conditional independence of the indicator items. This assumption follows from the idea that the latent variable is the common cause of the values of the indicator variables (see also Figure 1.1A). For example consider pessimistic thoughts, disturbed sleep, and irritability that are three symptoms of depression. The assumption states that these variables are related to each other, but only insofar that they are caused by the same disorder. All three take on the same value (present) when a respondent is depressed and the same value (absent) when not depressed. Conditional on (taking into account) depression there is no further relation between the three.

Of course this is quite an unrealistic assumption to fully meet. For example, a

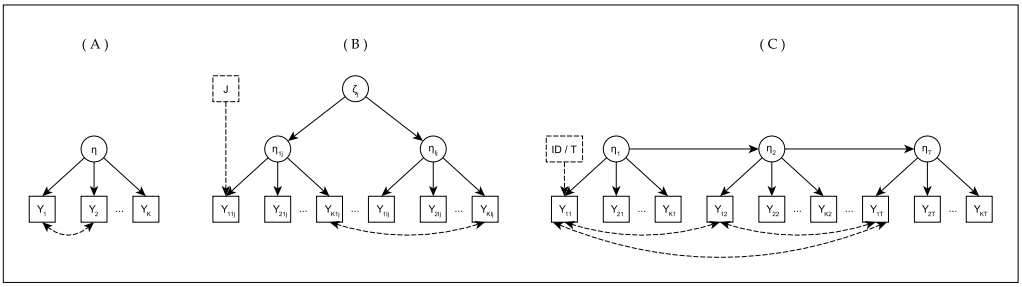

FIGURE 1.1: Overview of the (A) Latent class (B) Multilevel latent class (C) Latent Markov models. Dotted lines indicate conditional dependencies that are unwanted in common applications of these models.

life event such as moving house may cause irritability and disturbed sleep, but instead cause optimistic thoughts, implying that there will still be leftover association between irritability and disturbed sleep even after taking depression into account. The bivariate residual (BVR) (Vermunt & Magidson, 2016) is one way, amongst others (Glas, 1999; Asparouhov & Muthén, 2015) to quantify such leftover association. It does this by stating that, when everything is correct, the association between two variables predicted by the model should be the same as the association in the observed data. Because the data in these models is often wholly categorical, an efficient way to do this is by using Pearson's residual, and computing the difference between the predicted and the observed responses. Subsequently, there are several ways to determine whether this residual dependence between indicators is problematically large (Oberski, Van Kollenburg, & Vermunt, 2013; Khalid & Glas, 2016) or is likely to be due to random chance.

A similar assumption is made in the multilevel LC model. The multilevel extension explicitly exists to take into account the nested structure of the data, and to make sure that the systematic similarity of respondents that are a member of the same group is taken into account. Thus, conditional on the group-level latent variable, the group members' responses should be independent. For example schools can be classified as good, adequate, or bad in terms of academic performance by first classifying the students based on their grades and subsequently classifying schools by looking how many A-, B-, through F-grade students they have. The assumption is that the systematic similarity of students from the same school, or another type of observed group, is explained by the group-level latent variable, which in turn means that the entire list of grades, the response pattern, from one student should be independent of that list for any other student.

Until recently, neither this assumption nor the quality of the group-level classification in terms of how well the model fits each of the observed groups was quantifiable through a local fit statistic. The predominant, and very pragmatic, reason for this is that the interest in multilevel LC models has only picked up recently. With

that, the need for fit statistics that are easily obtainable and useful in an applied context has grown.

1.3 Goodness-of-Fit in the Latent Markov Model

In the LM model the dependence assumption is slightly different, in that here the data consists of multiple observations of the same person and often the main interest is the way in which respondents change over time. In terms of the initial classification of persons the same independence assumption holds as described in the previous section, namely that at each measurement occasion the indicator items should be independent given the latent variable. However, the within-respondent dependence is the direct substantive effect of interest (do respondents change between measurements and how), for which the probabilities should be estimated such that they not only describe change between occasions, but simultaneously capture the whole range of indirect associations with all other occasions (all the observations regardless of the distance between them should be conditionally independent, see also Figure 1.1 (B) and (C)).

This distinction between the estimated transitions and the indirect pattern of dependence is the result of combining a (first order) Markov chain and a latent measurement model. The Markov recursion states that there is only a direct effect between two adjacent measurement occasions. Thus, the current state (t) of a person is only affected by the previous one ($t-1$). This of course carries forward, if t is dependent on $t-1$ is dependent on $t-2$ there is a relation between t and $t-2$. However, this relation should be captured by the model without additional parameters.

The more constrained form of this model, where the assumptions of homogeneous transitions and measurement invariance are made, can intuitively be understood as estimating a classification for the very first measurement occasion, and one probability of moving to and from every state between each occasion. Here the transition probabilities describe the dependence between observations t and $t-1$ or $t+1$. Because of the forward recursion this is expected to indirectly model the relation between any pair of occasions, adjacent or very distant in time (Collins & Lanza, 2010).

This implies that in terms of the dependence structure, the same ideas as in the multilevel LC model play a role, but distinguishing between certain pairs of observations is of substantive interest. That is, all the observations nested in a respondent are expected to be conditionally independent, similar to all the respondents within a group. However, adjacent observations are of a different substantive interest. Moreover, the relation between measurements of the same variable (autocorrelation) is often found to be stronger in occasions that are closer together. These too may then be of more importance, or should be distinguished from, very distant occasions.

As for the substantive goals in the common applications of the model in the social sciences, it is important that the model correctly reproduces the item distributions at each occasion as well as for each respondent. That is, the response patterns of the

respondents should obviously be approximated, but so do the item distributions at each individual measurement occasion.

For this model no directly applicable local fit statistics are available that allow a simple test of the assumptions, causing much of the same problems as for the multilevel LC model. However, some statistics and ideas do exist for similar models. Most notably Titman (2007) takes a similar approach to quantifying misfit, but does so for the univariate model, keeping the residual statistics as more of a global fit indication. Furthermore, his approach is focused in particular on models where there is an all absorbing state, such as death of the respondent. Vasdekis, Cagnone, & Moustaki (2012) inspect univariate and bivariate residuals for longitudinal data, but do so per item by item and time by time pair, resulting in hundreds to thousands of residual statistics for only a moderate amount of items and measurement occasions.

1.4 Outline of the Thesis

In an effort to improve the applicability, ease of use, and especially correct use of the multilevel LC and the LM model this thesis aims to do the following: (a) Introduce local fit statistics for the multilevel LC model, (b) and for the LM model, whilst (c) assuring that the statistics are easily obtainable and (d) inspecting the power and type I error of the statistics to detect specific types of misfit.

The chapters in this thesis are, or are written as, journal articles and can be read separately and independently from each other. This does mean that there is some overlap in text, most notably in the sections that explain the technical details of the statistical models. Notation throughout the chapters is kept as consistently as possible, although some minor inconsistencies in the use of subscripts remain.

In Chapter 2 of the thesis the local fit statistics BVR-group and BVR-pair are introduced for the multilevel LC model, which are aimed at testing how well the model captures between-group differences and within-group similarities. Furthermore the bivariate residual is formulated such that it can be obtained for the multilevel model.

In Chapter 3 the properties of the BVR-group and BVR-pair statistics are studied for the multilevel LC model with an extensive simulation study to determine the power of these residuals to detect several types of misspecification.

In Chapter 4 five local fit statistics are proposed for the LM model, largely by adapting the BVR-group and BVR-pair for the multilevel model. The latter, which tests for residual within-respondent dependence, is furthermore split into a -lag1, -lag2 and general version to inspect residual dependence in adjacent, nearby and distant measurement occasions.

In Chapter 5 a new resampling approach is applied to the multilevel BVR statistics to see whether the required parametric bootstrap to obtain p-values can be sped up. This does require a slight adaptation of the residual statistics and leads to a fit indication that shows nuanced differences with the BVR statistics.

Chapter 2

Goodness-of-Fit of Multilevel Latent Class Models for Categorical Data

Abstract

In the context of multilevel latent class models, the goodness-of-fit depends on multiple aspects, among which are two local independence assumptions. However, because of the lack of local fit statistics, the model and any issues relating to model fit can only be inspected jointly through global fit statistics. This hinders the search for model improvements, as it cannot be determined where misfit originates and which of the many model adjustments may improve its fit. Also, when relying solely on global fit statistics, assumption violations may become obscured, leading to wrong substantive results. In this chapter, two local fit statistics are proposed to improve the understanding of the model, allow individual testing of the local independence assumptions, and inspect the fit of the higher level of the model. Through an application in which the local fit statistics group-variable residual and paired-case residual are used as guidance, it is shown that they pinpoint misfit, enhance the search for model improvements, provide substantive insight, and lead to a model with different substantive conclusions that would likely not have been found when relying on global information criteria. Both residuals can be obtained in the user-friendly LatentGOLD 5.0 software package.

This chapter is published as: Nagelkerke, E., Oberski, D. L. & Vermunt, J.K. (2016). Goodness-of-Fit of Multilevel Latent Class Models for Categorical Data. *Sociological Methodology, vol.46*(1), pp.252-282.

2.1 Introduction

Latent class (LC) analysis is mostly used to detect and develop a latent, or unobserved, classification of subjects based on multiple observed categorical characteristics. The usefulness of this application in many scientific fields combined with favorable properties, such as the ability to handle multiple dependent variables and measurement error, have recently caused a growing interest in LC analysis. This in turn has resulted in the development of several extensions to the regular model in an attempt to relax assumptions and make the method more widely applicable. An important extension that has gathered quite some attention is the multilevel LC model (Muthèn & Aspahourov, 2009; Vermunt, 2003, 2008).

Substantively the major benefit of this multilevel extension is that it allows simultaneous classification of groups and individuals. The regular LC model may either be used to distinguish typologies of the units under study that are systematically similar (Harrell et al., 2012), or find the most common characteristics of predetermined classes (Finch & Bronk, 2011; Laudy et al., 2005). The multilevel extension now makes it possible for nested categorical data in which a natural grouping is observed to also classify the groups based on the similarity of their members.

For example, employees can be classified in terms of job variety, which in turn is associated with job satisfaction and turnover intent (Lambert, Hogan, & Barton, 2001). However, the effect is likely to be moderated by the team context whereby correspondence rather than the absolute task variety is of importance. Perceiving far lower task variety compared to the team may cause diminished confidence and boredom, whereas far higher variety may induce stress. A simultaneous classification of both employees and the teams in which they are nested would allow the importance of this team context to be evaluated, providing more insight into outcomes such as frictional unemployment, employee burnout, or declines in overall job satisfaction.

In addition to the substantive application, the multilevel approach solves the statistical problem of dependent observations. Analogous to a multitude of statistical methods, LC analysis assumes that the units under study are independent of one another. However, this assumption does not hold when observing cases nested within a certain grouping, whether they are persons that belong to a particular group or repeated measures that belong to the same unit (Hox, 2010; Snijders & Bosker, 2012). In the example, the responses of employees from the same team cannot be assumed to be independent. An earlier solution to this dependence problem is the multiple-group approach (Clogg & Goodman, 1984), but it requires all parameters to be estimated separately for all groups, causing the method to lose its value when a large number of groups is observed.

Compared with the regular LC model, the multilevel LC model thus has additional substantive applications and offers a solution for categorical data in which there is dependence between observations. However, testing whether or not the

model is correctly specified and actually captures all the dependence is currently not possible in its own right, as inspecting model fit is limited to global tests, such as the chi-square (χ^2) or log-likelihood-ratio (L^2), and model comparisons through information criteria, such as the Bayesian information criterion (BIC) and Akaike information criterion (AIC). Although these tests and criteria can identify a well-fitting model, or the best fitting out of a series of alternative models, their global nature limits the control they provide. Especially when models become increasingly complex, the information available on the cause of better or worse fit becomes obscured. This in turn not only hinders the search for possible model improvements but also limits substantive understanding of the data.

To gain insight, understand the result of model adjustments, and detect specific misfit or violations of assumptions, these global criteria should ideally be supplemented with local fit statistics that single out and test one particular area of the model. In a regular LC model, such local fit measures exist in the form of the bivariate residual (BVR) (Vermunt & Magidson, 2005; see also Mavridis, Moustaki, & Knott, 2007) and a score-test approach that leads to modification indices (Glas, 1999; Oberski, Van Kollenburg, & Vermunt, 2013). Both test the local independence assumption that is central to the LC model and evaluate the degree to which the model captures the association between all pairs of observed variables. As such, these measures indicate why one model fits better or worse, pinpoint violations of the local independence assumption, and facilitate the search for model improvements. For the multilevel LC model, however, there are currently no local fit statistics that give these insights on the group level.

Here we propose two complementary diagnostic measures that enhance exactly these abilities to detect a particular type of model misfit and increase the understanding of the fitted model for multilevel LC analysis. Both take the form of a Pearson residual and relate to the higher level of a multilevel LC model. The first residual, BVR-group, relates to the item distributions and is considered a between-group measure. It can be used to evaluate the difference in responses between groups and to detect misfit that originates from the model not fitting particular groups as well as others. The second residual, BVR-pair, is a within-group measure in the sense that it can be used to evaluate the degree of similarity among cases within a group, and it is indicative of misfit that originates from any leftover dependence among the units within groups.

The remainder of this chapter is structured as follows. In section 2, we introduce the multilevel LC model. In section 3, we discuss the problems with model fit statistics more elaborately, as well as the existing BVR, and introduce the proposed residuals. In section 4, we demonstrate the use of the residuals as local fit measures and the way in which they may affect substantive conclusions by applying them to the job variability data used by Vermunt (2003). In section 5, we use a small simulation study to demonstrate that the proposed measures have adequate power and type I error.

2.2 The Multilevel Latent Class Model

The multilevel LC model can be expressed using two equations: one for the lower level denoting the conditional probability of all responses given by a unit and one for the higher-level marginal probability of all response patterns per group (Vermunt, 2003; Lukočiené, Varriale, & Vermunt, 2010). The expression for the lower level is essentially that of an LC model, but in the case of a multilevel structure it is made conditional on the LC membership of the group (Vermunt, 2003, 2008).

Let the response of individual i in group j on item k be denoted as y_{ijk}, with a total of J groups, each having n_j individual members summing to N, and a total of K items, each having R_k categories. All responses to the K items of person i in group j are denoted as the vector $\mathbf{y}_{ij}$, with $\mathbf{r}$ referring to one particular answer pattern of length K when no values are missing and r_k referring to a particular response to item k. The latent variable η_{ij} that classifies the units within groups has C latent classes and the latent variable ζ_j that classifies the groups has G latent classes, with c and g referring to one of these classes. Assuming conditional independence, the lower level of the multilevel LC model is expressed as

$$P(\mathbf{y}_{ij} = \mathbf{r}|\zeta_j = g) = \sum_{c=1}^{C} P(\eta_{ij} = c|\zeta_j = g) \prod_{k=1}^{K} P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g). \quad (2.1)$$

Removing the conditioning on the group-level latent variable (ζ_j) from Equation 2.1 results in the standard LC model, in which the probability of observing a particular response pattern $\mathbf{r}$ is a combination of the prevalence of LC c on the latent variable η_{ij} and the probabilities of observing the combination of the responses r_k conditional on the unit's class membership. In the multilevel LC model, all these terms are made conditional on the LC membership of the group a unit belongs to ($\zeta_j = g$), such that groups can be classified along G LCs and the probability of an individual response pattern is affected by the group-level class membership.

The expression for the higher level of the model then denotes the marginal probability of all response patterns of individuals within group j as $\mathbf{y}_j$, with $\mathbf{s}$ denoting a particular combination of response patterns of length $n_j * K$. Here an assumption of independence is required as well, but now the full response patterns of individuals rather than the responses to one item should be independent:

$$P(\mathbf{y}_j = \mathbf{s}) = \sum_{g=1}^{G} P(\zeta_j = g) \prod_{i=1}^{n_j} P(\mathbf{y}_{ij} = \mathbf{r}|\zeta_j = g) \quad (2.2)$$

The probability of observing the vector $\mathbf{y}_j$ of all individual response patterns $\mathbf{s}$ in group j is a combination of the prevalence, or size, of a particular group-level LC g on the latent variable ζ_j and the probabilities of observing the combination of the individual answer patterns $\mathbf{r}$ conditional on the LC membership of the group.

It should be noted that these two expressions result in a model in which both the

lower-level class prevalence and the response probabilities can differ between all higher-level classes. Although a multitude of constraints is possible, two are most commonly used in practice, the first of which leads to the most used model that simultaneously classifies higher- and lower-level units. The first constraint $P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) = P(y_{ijk} = r_k|\eta_{ij} = c)$ causes the response probabilities on the lower level to be independent of the higher-level class membership but the class sizes to be estimated freely (Lukočienė, Varriale, & Vermunt, 2010; Vermunt, 2003, 2008). The second possibility is to constrain the model by setting $P(\eta_{ij} = c|\zeta_j = g) = P(\eta_{ij} = c)$, causing the response probabilities to be estimated freely but the lower-level class membership to be independent of higher-level class membership (Vermunt, 2004; Lukočienė, Varriale, & Vermunt, 2010).

2.3 Goodness-of-Fit

In this multilevel LC model, there are several key issues relating to model fit. There are the two central assumptions —namely, the local independence of item responses on the lower level and the conditional independence of response patterns of individuals on the higher level— and there are the goals of correctly reproducing the item distributions or observed frequencies for both the individual observations as well as for the groups. These latter goals relate to arriving at a correct classification on both levels and to obtaining the conditional probabilities of interest depending on the substantive goal and specification of the model (Goodman, 2002).

Improving the fit of this model can be achieved in a multitude of ways that improve the quality of the prediction, or relax an assumption. An LC or group-level LC can be added, for example. Or, when keeping the same number of classes, a covariance between any combination of observed variables may be modeled, as well as any direct effect from the group-level latent variable to an observed variable. Although it is also possible to add additional categorical or continuous latent variables to the model, for conciseness, these options are not explored in the application.

Unfortunately, despite these different sources of misfit and the many ways to adjust the model, there is little information available as to where model misfit originates and what the effects are of model adjustments. Currently only the local independence assumption on the lower level of the model —the independence of responses conditional on the latent variable— can be inspected through the BVR. The analogous assumption on the higher level —the independence of response patterns conditional on the group-level latent variable— the quality with which the model describes the individual responses, and the degree to which the model correctly describes the groups can only be assessed jointly through global statistics. That is, the fit of the model as a whole is considered, rather than any of the individual aspects of the model.

As a result local misfit may go unnoticed, because even when a model shows adequate global fit, it may still be misspecified. In such cases, a type of local misfit

averages out with other, correctly specified, areas of the model. This problem is reinforced when using information criteria, such as the BIC and the AIC, which only compare estimated models. As long as all estimated models in such cases violate one or more assumptions, selecting the best one will still result in using a model that does not fit the data correctly. Ultimately this may lead to a wrong classification and wrong substantive conclusions, especially when the classification is used in subsequent analyses to relate classes to outcomes.

Of course, these problems with global fit measures apply to almost all statistical methods, but they do become more pressing in complex models as the possible sources of misfit are abundant. This is especially clear in multilevel models, for which both levels are considered simultaneously. For multilevel structural equation modeling, several solutions have been offered to evaluate the fit separately for different levels. Yuan & Bentler (2007) did so by estimating the saturated covariance matrices for each level of the full model and subsequently treat these as observed single-level data to test the hypothesized model one level at a time. As such they obtained common fit indices for each level individually. Ryu & West (2009) developed a similar approach whereby the model is initially estimated as hypothesized and subsequently reestimated several times, each time saturating one of the levels.

Although both are elegant solutions to localize model misfit, such methods do not apply to LC analysis, as the higher level cannot be estimated independently from the lower level. As was shown in Equation 2.2, the vector of group-level patterns is directly related to the estimated answer patterns for respondents. When the lower level is saturated, this also greatly improves the fit on the higher level of the model. Furthermore, even though these methods are able to separate the misfit on different levels, they still are not local fit statistics in the true sense that they are able to pinpoint the assumption violation, misspecification, or variable that causes the misfit. That is, even when the level at which misfit occurs can be determined, the possibilities to improve the model remain plentiful and require more precise measures to be detected.

To address this problem, two local fit statistics for multilevel LC models are proposed in the sections that follow, which aim to test specific areas of the model individually. The first tests the reproduction of univariate item distributions in all the groups and provides a partial test of how well the higher level of the model fits the data. The second is aimed at testing the conditional independence of response patterns and in combination with the BVR allows a test of two central assumptions of the model. Both provide information on the location and extent of misfit.

2.3.1 Bivariate Residual (BVR)

To show how the proposed statistics fit the LC framework, and for the sake of completeness, the existing BVR is briefly introduced. Vermunt & Magidson (2013) constructed the BVR to test the assumption of local independence for all pairs of observed variables in a regular LC model, but the test can be applied identically to the

lower level of a multilevel model. The BVR assesses the difference between the observed frequencies ($n_{rr'}$) and the model expected frequencies ($m_{rr'}$) in the two-way cross-tabulation of items k and k' by a Pearson statistic divided by its number of degrees of freedom (see also Vasdekis, Cagnone, & Moustaki, 2012; Bartholomew & Leung, 2002); that is,

$$BVR_{kk'} = \frac{1}{(R_k - 1)(R'_k - 1)} \sum_{r=1}^{R_k} \sum_{r'=1}^{R_{k'}} \frac{(n_{rr'} - m_{rr'})^2}{m_{rr'}}. \tag{2.3}$$

The expected frequencies follow from the LC model, which assumes conditional independence of item responses given LC membership. More specifically, they are obtained by multiplying the class-specific probabilities of the response r on item k and response r' on item k' and summing these over the LCs using the class membership probabilities as weight. For an LC model without a multilevel structure, we obtain

$$m_{rr'} = \sum_{i=1}^{N} \sum_{c=1}^{C} P(y_{ik} = r_k|\eta_i = c)P(y_{ik'} = r_{k'}|\eta_i = c)P(\eta_i = c|\mathbf{y}_i = \mathbf{r}). \tag{2.4}$$

When no values are missing, the same $m_{rr'}$ can be obtained by using $P(\eta_i = c)$ as a weight instead of $P(\eta_i = c|\mathbf{y}_i = \mathbf{r})$ and multiplying the sum over classes by N rather than summing it over N, because $P(\eta_i = c)$ equals the average $P(\eta_i = c|\mathbf{y}_i = \mathbf{r})$ for the complete sample. However, in the case of missing values, the observed frequencies contain only those cases for which both variables are observed. To obtain the corresponding expected frequencies, the class membership probabilities should be based on this subsample. That is, using $P(\eta_i = c)$ is not appropriate, and the frequency should be obtained by summing over the cases with both variables observed, using $P(\eta_i = c|\mathbf{y}_i = \mathbf{r})$ as a weight.

The above formulation for $m_{rr'}$ can be easily generalized to be applicable in a multilevel LC analysis. The sum over LCs must then contain the joint posterior probability of the lower- and higher-level latent variables, and the sum over individuals must be over both groups and individuals within groups:

$$m_{rr'} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{g=1}^{G} \sum_{c=1}^{C} P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) \\ P(y_{ijk'} = r_{k'}|\eta_{ij} = c, \zeta_j = g)P(\eta_{ij} = c, \zeta_j = g|\mathbf{y}_j = \mathbf{s}). \tag{2.5}$$

Any deviation between the observed and the predicted frequency, which assumes local independence of items given LC membership, is now contained in the residual.

2.3.2 Group-variable Residual (BVR-group)

To further deconstruct global misfit, we here propose a group-variable residual, BVR-group. As was shown in Equation 2.2 the response vector $\mathbf{y}_j$ containing all individual response patterns is a function of the size of the group-level class and the individual answer patterns. This implies that, among other things, the univariate response frequencies within each group should be modeled correctly for the LC solution to be correct. Because the observed group membership can be understood as a nominal covariate in a multilevel LC model, the BVR can be adapted to assess the response to a nominal dependent variable and group membership:

$$BVR_{group.k} = \frac{1}{(R_k - 1)(J - 1)} \sum_{j=1}^{J} \sum_{r=1}^{R_k} \frac{(n_{jr} - m_{jr})^2}{m_{jr}} \tag{2.6}$$

The observed frequency n_{jr} here is simply the number of units in group j with response r_k. The expected frequencies m_{jr} can be obtained from the individual probabilities $P(y_{ijk} = r_k)$:

$$P(y_{ijk} = r_k) = \sum_{g=1}^{G} P(y_{ijk} = r_k|\zeta_j = g)P(\zeta_j = g|\mathbf{y}_j = \mathbf{s}), \tag{2.7}$$

where

$$P(y_{ijk} = r_k|\zeta_j = g) = \sum_{c=1}^{C} P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g)P(\eta_{ij} = c|\zeta_j = g). \tag{2.8}$$

Then

$$m_{jr} = \sum_{i=1}^{n_j} P(y_{ijk} = r_k) = \sum_{i=1}^{n_j} \sum_{g=1}^{G} P(y_{ijk} = r_k|\zeta_j = g)P(\zeta_j = g|\mathbf{y}_j = \mathbf{s}). \tag{2.9}$$

Thus, the probability of a particular response is summed over all group members to obtain its frequency within the group, and it is itself a function of the group-class response probabilities and the group-class membership probabilities. It should be noted that for the class membership on the group level, the posterior probability $P(\zeta_j = g|\mathbf{y}_j = \mathbf{s})$ is used. Because the interest lies in testing the group by variable relationships and aggregating these over the groups, all available information on the groups should be used, as contained in the posterior.

The statistic itself is computed for all groups separately and summed over the groups to test the assumption of correct model fit in each of the groups. This sum is additionally divided by $(R_k-1)(J-1)$. The BVR-group now equals the average contribution to the residual per degree of freedom. That is, the dimension of the matrix to which Equation 2.6 is applied is $R_k \times J$, resulting in $(R_k-1)(J-1)$ nonredundant parameters. Correcting for both R_k and J standardizes the BVR-group such that it is not affected by the number of groups or the number of categories on the variable.

As can be seen in Equations 2.7 through 2.9, a special case exists when the nested structure of the data is ignored by estimating the multilevel LC model with only one group-level class. The results are identical to omitting the group-level latent variable altogether and ensures that the BVR-group is independent from the number of lower-level classes to obtain its baseline value, which is substantively indicative of the between-group heterogeneity or the between-group variance. For this model, the residual is then broadly comparable to the empirical Bayes estimates as used in linear multilevel models. Although their common use is to test the normality assumption on the higher level, they can also be used to construct influence diagnostics (Snijders & Berkhof, 2008) and as such are indicative of misfit.

2.3.3 Paired-case Residual (BVR-pair)

In a multilevel LC model, the higher level has a local independence assumption similar to that of the lower level. Where the assumption in Equation 2.1 is that the responses r_k are independent for all the K items per individual, in Equation 2.2 the response patterns $\mathbf{r}$ are assumed independent for all the individuals per group. However, to capture this dependence among units within a group, the responses of the individual members should be related to one another. This cannot be done as straightforwardly as is the case for the dependence between item pairs. Where the response frequencies for the latter can be cross-tabulated directly, the cross-tabulation of dependence among units requires all units within a group to be related. An intuitive approach to do so is to create all pairs of units within every group and obtain the pairwise response frequencies. The expected and observed response frequencies can then be compared again:

$$BVR_{pair} = \frac{J}{N}\frac{1}{R_k(R_k-1)/2} \left[\sum_{r}^{R_k}\sum_{r'>r}^{R_{k'}} \frac{((n_{krr'}+n_{kr'r})-(m_{krr'}+m_{kr'r}))^2}{m_{krr'}+m_{kr'r}} + \sum_{r}\frac{(n_{krr}-m_{krr})^2}{m_{krr}}\right]. \quad (2.10)$$

To illustrate, consider a group containing five observations, with responses to one of multiple variables, as in Table 2.1. The residual can be understood as considering the combined responses r and r' of cases i and i' to item k as one element. To obtain the observed frequencies, a square contingency table of which the order is equal to the number of categories on the variable of interest can then be made per pair. The cell that identifies the actual answer pattern of that pair of cases has a frequency of one and all else equals zero.

The corresponding predicted probability of a certain pair of responses follows from the combined probability of person i giving response r and person i' giving response r' conditional on the group-level class:

TABLE 2.1: Obtaining the observed pairwise response frequencies

Data

Obs	Var	Group
A	0	1
B	0	1
C	1	1
D	0	1
E	1	1
F	0	2
G	1	2
H	...	...

	B	
	0	1
A 0	1	0
1	0	0

	C	
	0	1
A 0	0	1
1	0	0

	D	
	0	1
A 0	1	0
1	0	0

	E	
	0	1
A 0	0	1
1	0	0

	C	
	0	1
B 0	0	1
1	0	0

	D	
	0	1
B 0	1	0
1	0	0

	E	
	0	1
B 0	0	1
1	0	0

	D	
	0	1
C 0	0	0
1	1	0

	E	
	0	1
C 0	0	0
1	0	1

	E	
	0	1
D 0	0	1
0	0	0

$$P(y_{ijk} = r_k, y_{i'jk} = r'_k) = \sum_{g=1}^{G} P(y_{ijk} = r_k|\zeta_j = g)P(y_{i'jk} = r'_k)P(\zeta_j = g|\mathbf{y}_j = \mathbf{s}), \quad (2.11)$$

where $P(y_{ijk} = r_k|\zeta_j = g)$ can be obtained by Equation 2.8. Because these probabilities are only conditional on the group-level latent variable in a model without covariates, they are identical for identical patterns, and the order of the responses is interchangeable. That is, within a group only the probabilities on the diagonal and either the upper or lower off-diagonal need to be obtained. Aggregating these probabilities to arrive at the expected frequencies can then be done by multiplying the probability of a pair with the number of pairs $n_j(n_j - 1)/2$:

$$m_{krr'} = \sum_{j=1}^{J} (n_j(n_j - 1)/2)P(y_{ijk} = r_{k'}, y_{i'jk} = r'_k). \quad (2.12)$$

Again, as is done for the BVR-group, the posterior probability is used in Equation 2.11 to obtain this estimated frequency. In this case the main reason is that this weighting is more appropriate in cases in which groups are of different sizes and thus contain different numbers of pairs per group. As can be seen from Equations 2.10 and 2.12, in comparison with Equations 2.6 and 2.9, the BVR-pair is not obtained for each group separately and is only subsequently summed over the groups, but the aggregation already occurs when computing the expected frequencies. By weighting according to the posterior probability $P(\zeta_j = g|\mathbf{y}_j = \mathbf{s})$, the expected frequencies account in the best manner for unequal group sizes. With equal group sizes, using posterior or unconditional class membership probabilities will give the same expected frequencies.

The observed frequency of pairs can now be obtained by summing the pairwise tables from Table 2.1, as is done in Table 2.2. The probability of a pair follows from equation 2.11 and the expected frequency from Equation 2.12. For the illustration, the probabilities from the first model in the application section are used.

Here, the structure of equation 2.10 also becomes clear. Note that because the

TABLE 2.2: Obtaining the pairwise residual contribution per answer pattern

	Observed				Probability				Expected				Residual Contr.			
		i'				r'				i'				i'		
		0	1			0	1			0	1			0	1
i	0	3	5	r	0	.415	.225	i	0	4.152	2.249	i	0	.320	.056
	1	1	1		1	.225	.135		1	2.249	1.351		1	-	.091

$$BVR_{pair} = \frac{1}{6}\frac{1}{2(2-1)/2}(.320 + .056 + .091) = 0.079$$

order of the observations within a group is arbitrary, observing a 0-1 pair is in fact the same as observing a 1-0 pair. This is why the symmetric off-diagonal elements of the table are combined in the first summation in equation 2.10. The latter part of equation 2.10 adds the discrepancy between the observed and expected frequencies on the diagonal.

To finally arrive at the BVR-pair the resulting residual is divided in such a way that the statistic equals the contribution to the residual per degree of freedom, in this case $R_k(R_k - 1)/2$ given the symmetry on the off-diagonals. In addition, the raw residual is divided by the average group size to avoid extremely large values, which are likely to occur because the theoretical maximum value of the statistic increases as a triangular sequence with n_j.

Unfortunately, the univariate marginal values for the resulting tables are not reproduced correctly when groups differ in size, in which case $(n_{rr'} + n_{r'r}) \neq (m_{rr'} + m_{r'r})$, which is also the case in the illustration. The cause is simply that an observation in a larger group is in more pairs than an observation in a smaller group. Differences between the observed (n) and expected (m) frequencies would then not only reflect the degree to which the model captures dependence between cases, but the residual would also partly reflect the difference in the univariate distribution. This changes the interpretation of the BVR-pair which is unnecessary because the univariate distributions are always correctly reproduced by the model.

Therefore, a number of iterative proportional fitting (IPF) cycles are used to equate the reproduced and observed marginal frequencies and reduce the BVR-pair to zero when there is no residual dependence. The pairwise contingency table is made symmetrical first, such that answer patterns that differ only in respect to the order of the responses have the same frequency. As mentioned, the probability and thus the expected frequency of a certain pair of responses are identical regardless of order, but this is not necessarily the way in which they are observed.

In the IPF procedure the cells in the expected frequency table are adjusted so that its marginals match the observed marginals. The subsequent iterations alternate between row and column adjustments where each cell is multiplied by the ratio

TABLE 2.3: Iterative Proportional Fitting (IPF) Illustration

Observed				
		i'		
		0	1	
i	0	3	(5+1)/2	6
	1	(5+1)/2	1	4
		6	4	10

Expected				
		i'		
		0	1	
i	0	4.152	2.249	6.401
	1	2.249	1.351	3.599
		6.401	3.599	10

IPF Cycle 1 - Row				
		i'		
		0	1	
i	0	3.892	2.108	6
	1	2.499	1.501	4
		6.391	3.609	10

IPF Cycle 1 - Column				
		i'		
		0	1	
i	0	3.654	2.236	5.99
	1	2.346	1.664	4.01
		6	*4*	10

IPF Cycle 2 - Row				
		i'		
		0	1	
i	0	3.66	2.34	*6*
	1	2.34	1.66	*4*
		6	3.999	10

IPF Cycle 2 - Column				
		i'		
		0	1	
i	0	3.66	2.34	6
	1	2.34	1.66	4
		6	*4*	10

Row operation: Multiply cell by the ratio between the observed and expected row marginals: cell(observed row/expected row).
Column operation: Multiply cell by the ratio between the observed and expected column marginals: cell(observed column/expected column).

between the observed and the expected row (column) marginal. This process converges to a table with marginals equal to the observed marginal frequencies while retaining the cross-product ratios within the table (Bishop, Fienberg, & Holland, 1975). An example can be found in Table 2.3.

The resulting BVR-pair statistic reduces to zero when the model captures all the dependence among cases within a group. Identical to the BVR-group its baseline value can be obtained by estimating the model where the nesting of the data is ignored by modeling only one group-level class. The statistic is broadly comparable with the residual intraclass correlation in mixed models, which is the degree of dependence that is not captured by the model when controlling for the independent variables (Snijders & Bosker, 2012). The BVR-pair is similarly related to the uncaptured dependence and indicative of the homogeneity within groups that is ignored when the nested structure of the data is not or is only partially reproduced.

2.3.4 Bootstrap

The BVR, BVR-group and BVR-pair residuals are all obtained identically to Pearson residuals. However, for the BVR, it is known that it does not follow the chi-square distribution, and the same is expected to be true for the two proposed measures. To still obtain p-values for the residuals, a parametric bootstrap can be used (Langeheine, Pannekoek, & Van de Pol, 1996), which is known to work for the BVR (Oberski, Van Kollenburg, & Vermunt, 2013). On the basis of the maximum likelihood estimate, the bootstrap in this instance samples group-class membership, class membership conditional on group-class membership, and the responses conditional

on the membership of both. This results in alternative data sets with the same structure as the original to which the model is fitted. For each of these refitted models, the BVR values are obtained. The estimated p-value then is the proportion of replicated models in which the BVR residuals are larger than in the original model (Vermunt & Magidson, 2013). As such the BVR-group and BVR-pair are compared not with an asymptotic distribution but rather with an empirical distribution constructed by simulation. The bootstrap p-values can be used for hypothesis testing, that is, for determining whether potential assumption violations are statistically significant.

2.4 Application: Improving the Job Variety Classification

To illustrate the usefulness of the BVR-pair and BVR-group we apply them here to a data example in which both employees and the teams in which they are nested are classified on the basis of task variety. This is one of the examples Vermunt (2003) used when introducing multilevel LC analysis, which provides the opportunity to see whether the original solution can be improved on the basis of the two residuals.

The variety in the tasks of employees, as well as the degree to which they feel that their capacities are put to good use, has been found to affect job satisfaction and turnover intent (Lambert, Hogan, & Barton, 2001; Fila, Paik, Griffeth, & Allen, 2014). Although these outcomes are inherently individual, the broader context of the team, department, or organization plays an important role in shaping these effects. Gunter & Furnham (1996), for example, found that job variety has an opposite effect on job satisfaction in two different organizations, and Van Mierlo, Rutte, Kompier, & Doorewaard (2005) gave a broad overview of studies in which individual and team tasks affected several outcomes.

One of the ways in which context may affect job satisfaction and turnover intent may be through peer perceptions (e.g. Liu, Mitchell, Lee, Holtom, & Hinkin, 2012). When direct coworkers perceive their jobs as highly varied when individuals do not, this may adversely affect job satisfaction. In contrast, teams with larger differences in task variety may be better able to distribute the work, improving individual job satisfaction and reducing turnover intent. By obtaining a classification of teams through multilevel LC analysis on the basis of the perceived job variety classification of the employees, it becomes possible to detect such differences in team composition and investigate these questions.

Although relating the classification to an outcome variable is beyond the scope of this example, the use of simultaneous classification can be easily extended. For example, when job design is aggregated, it may explain frictional unemployment caused by a mismatch between companies and the workforce in a region, a classification of countries on the basis of the degree of religiosity of their populations may form an explanation for policy differences, or a classification of pupils and their groups may be used to enhance school class composition (see also Bennink, Croon, Keuning, & Vermunt, 2014).

TABLE 2.4: BIC values for 29 models assuming local independence of items and indirect effects of the group-level latent variable

	Lower-level Classes				
Group-level Classes	2	3	4	5	6
1	4,820	4,818	4,837	4,861	_[c]
2	4,786	***4,785***	4,799	4,482	4,844
3	4,794	4,795	4,794	4,814	4,837
4	4,802	4,806	4,808	4,826	4,850
5	4,811	4,818	4,822	4,839	4,865
6	4,820	4,831	4,838	4,857	4,881

[a] Values obtained using the number of groups J as the sample size in the BIC computation.
[b] Constraint: $P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) = P(y_{ijk} = r_k|\eta_{ij} = c)$.
[c] Unidentified.

However, when the LC model is incorrectly specified or violates assumptions, there is a possibility not only that teams and employees may be wrongly classified but also that the relationship between an outcome and the classification may be similarly unsound. This first step of classification is clearly an important one, because a wrong classification may result in wrong substantive conclusions on the actual goal of the study. Here the classification is reexamined using the proposed BVR-group and BVR-pair statistics to demonstrate their use. After excluding all cases with missing values and two teams with only one member, the data contain 848 cases in 86 teams and are similar to the data used by Vermunt (2003) and Vermunt & Magidson (2005), as collected by Van Mierlo (2003). For all employees, the perception of task variety in their jobs was measured with five categorical items, of which the four categories are collapsed to make them dichotomous. The variable measuring task repetitiveness is coded inversely with the other variables, such that a higher score reflects lower repetition and all scores are substantively in accordance. All models are estimated in LatentGOLD 5.0. The LatentGOLD syntax and survey wording are provided in Appendix A and Appendix B, respectively. The data set itself is included in LatentGOLD as example data.

Because the BIC is currently the main criterion for model selection, selecting the best fitting from a series of alternative models, Table 2.4 depicts the BIC values for 29 models with differing numbers of classes. All of these models assume conditional independence between the five items, contain one latent variable on both levels (η and ζ), and allow only an indirect effect of the group-level latent variable ζ on the items through the lower-level latent variable η (see also Vermunt, 2003). It should be noted that these BIC values are computed using the number of groups as the sample size, rather than the number of cases, as this is found to be the more appropriate sample size to determine the number of classes in multilevel LC models (Lukočienė et al., 2010; Lukočienė & Vermunt, 2010).

On the basis of these values, the model with two group-level and three low-level classes would be the best fitting, resulting in the profile depicted in Table 2.5. On the lower level, the largest of the three classes is one in which people report high

TABLE 2.5: Latent class profile of the two class, three group-level class model

	Group class 1 Diverse	Group-class 2 Uniform	Class 1 Diverse	Class 2 Structure	Class 3 Creative	Overall
Nonrepetitive	.428	.279	.515	.125	.225	.385
Creative	.631	.382	.707	.065	.914	.558
Diverse	.792	.480	.961	.146	.483	.700
Capacity	.730	.578	.837	.439	.350	.685
Variation	.754	.461	.964	.192	.000[a]	.668
Class 1	.752	.371				
Class 2	.150	.537				
Class 3	.098	.092				
Prevalence	.707	.293	.640	.263	.097	

[a] Boundary solution.

levels of task variation and creativity. The second class is one in which people report having repetitive, uncreative, and unvaried tasks. The third is a class with highly creative tasks, yet quite unvaried and repetitive. On the group level, the classes are less distinguished in their overall profile. Members of teams in the first group-level class are most likely to belong to the first individual-level class and those of the second higher-level class to the second lower-level class. Overall then the team profile of the first group-level class is mostly that of diverse, varied, and challenging tasks, whereas the second class has more repetitive tasks that allow less creativity.

However, the two problems laid out in section 2.3 would arise when this model would be accepted solely on the basis of the BIC value. First, the BIC identifies the best alternative out of the models presented, but it does not guarantee that no assumptions are violated, that is, that the model picks up all relevant aspects in the data. If this is not the case, the classification described in Table 2.5 could be faulty, and any further analysis to relate this classification to outcomes may also be affected negatively. Second, many alternative models can be specified, other than those with differing numbers of classes.

In all the estimated models, conditional independence of the observed items is assumed, which can be relaxed by allowing one or more covariances between the observed variables. Furthermore, the effect of the group-level LC on the observed variables is assumed to be fully mediated by lower-level class membership. This too can be relaxed by allowing direct effects from the higher-level latent variable on any of the items. The prohibitive difficulty of improving the model through trial and error, or even considering the option of estimating all possible models, now quickly becomes clear. When keeping the number of classes constant, there are 1,024 different combinations of allowable covariances and, for each of these combinations, another 32 possible combinations of direct effects. If the possibility of equating certain parameters to one another is also considered, this model can be adjusted in 17 factorial different ways.

TABLE 2.6: BVR, BVR-group, and BVR-pair residuals for the three class, two group-level class model. Bootstrap p-values between parentheses

	Nonrepetitive	Creative	Diverse	Capacities	Variation
Creative	0.763 (.242)				
Diverse	0.248 (.282)	0.028 (.442)			
Capacities	0.183 (.570)	0.359 (.308)	0.504 (.106)		
Variation	0.010 (.706)	0.036 (.272)	0.153 (.016)	0.011 (.790)	
BVR-group	1.586 (.000)	1.051 (.000)	0.788 (.164)	1.072 (.132)	0.816 (.316)
BVR-pair	1.740 (.000)	0.570 (.028)	0.123 (.296)	0.366 (.098)	0.000 (.974)

Note: Bayesian information criterion = 4,785.3.

To illustrate how the local fit measures may largely resolve the problem of identifying misfit without the need to estimate many additional models, the residual measures for the model with the lowest BIC are presented in Table 2.6 with bootstrapped p-values for all BVR measures in parentheses. The regular BVR indicates that the variable measuring the diversity of a person's job shows some residual dependence with the variable measuring job variation, which substantively should come as no surprise. On the higher level, the BVR-group and BVR-pair also show assumption violations, whereby the repetitive and creative variables both show dependence between cases that is not captured by the model, as well as an incorrectly reproduced item distribution between the groups. So, even though it is the best alternative out of 30 models, the three individual-level, two group-level class model violates the three tested assumptions to some extent.

From Table 2.4, it can be concluded that improving this model is not achieved by increasing the number of classes. Inspecting the BVR measures for these models leads to the same conclusion, as a combination of problems on both levels of the model persists when increasing either the number of classes on the lower level, the higher level, or both.

Thus, to improve this model, a solution other than increasing the number of classes is required. Starting model improvements on the lower level of the model is often the most fruitful, as it is more likely that group-level dependence is introduced by having a wrong specification on the lower level than the reverse (Lukočiené et al., 2010). This is due to the higher level classification being partly determined by the classes on the lower level, as can be seen in equation 2.2.

Substantively, the significant dependence between the self-reported variation and diversity of work is sensible, and including a covariance between these two variables seems justified. As shown in Table 2.7, adding this covariance removes any problematic bivariate dependence on the lower level of the model.

Considering the BVR-group and BVR-pair statistics, the logical next step is to add a direct effect from the group-level latent variable on the repetitive variable. Such a direct effect is the most parsimonious solution in an attempt to capture more dependence and improve within-group model fit regarding the repetitive variable, adding only one parameter. Substantively too, there is evidence that the differences

TABLE 2.7: Residuals for the three class, two group-level class model. Covariance between Variation and Diverse. Bootstrap p-values between parentheses

	Nonrepetitive	Creative	Diverse	Capacities	Variation
Creative	0.101 (.642)				
Diverse	0.602 (.104)	0.022 (.514)			
Capacities	0.871 (.184)	0.001 (.938)	0.178 (.264)		
Variation	0.062 (.400)	0.042 (.316)	0.000 (.999)	0.028 (.670)	
BVR-group	1.576 (.000)	0.973 (.140)	0.776 (.264)	1.037 (.194)	0.842 (.312)
BVR-pair	1.523 (.000)	0.294 (.130)	0.128 (.296)	0.256 (.138)	0.011 (.780)

Note: Bayesian information criterion = 4,783.2.

in repetitive work between teams reflect on that of the individual tasks (Van Mierlo, 2003).

After adding this effect, problems arise in all five variables, as depicted in Table 2.8, causing the model to no longer describe the within-team item distributions correctly; nor does it adequately capture the dependence between cases. Yet despite the large shift on the group level of the model, the lower level does not show any problems. The interpretations of the individual-level classes (not reported) also do not change, indicating that the problems are largely the result of a failure to capture team differences correctly. Given that there are problems with all five variables on the group level of the model, adding an additional group-level class is the best option here.

Adding a third group-level class indeed solves most problems on the higher level of the model, as can be seen from Table 2.9. In this model, the covariance between the variation and diverse variable, as well as the direct effect on the repetitive variable, is retained. As a final adaptation, a direct effect from the group-level latent variable on the creative variable is added, following the BVR-group value, and the reasoning that the structure of a team and the overall packet of tasks it realizes may have a direct effect on the creativity an employee has in accomplishing their share of the teamwork.

In Table 2.10, the BVR, BVR-group and BVR-pair residuals for the final model are presented. Further attempts to make this model more parsimonious result in models in which significant residuals are reintroduced. Note that the model chosen has a higher BIC value than the previous model (4,768.9 compared with 4,775.3), but given the focus on model fit and misfit, we opt for the less parsimonious model. This choice depends on the goal of the model specification. If the goal is to obtain high posterior probabilities, the model for which the residuals are presented in Table 2.9 would be preferred (Burnham & Anderson, 2002; Hamaker, Van Hattum, Kuiper, & Hoijtink, 2011).

The profile of this final model is presented in Table 2.11. Comparing these results with those in Table 2.5, it becomes clear that the individual-level classification is practically identical to that obtained in the model with two group-level LCs and three individual-level LCs. On the group level, the additions to the model, an extra

TABLE 2.8: Residuals for the three class, two group-level class model. Covariance between Variation and Diverse and direct effect from the group-level latent variable on Nonrepetitive. Bootstrap p-values between parentheses

	Nonrepetitive	Creative	Diverse	Capacities	Variation
Creative	0.004 (.922)				
Diverse	0.737 (.082)	0.068 (.204)			
Capacities	0.962 (.180)	0.026 (.732)	0.046 (.670)		
Variation	0.019 (.664)	0.034 (.212)	0.000 (.999)	0.090 (.432)	
BVR-group	1.544 (.000)	1.405 (.000)	1.356 (.000)	1.194 (.040)	1.125 (.048)
BVR-pair	1.657 (.000)	0.930 (.006)	1.325 (.002)	0.458 (.048)	0.280 (.070)

Note: Bayesian information criterion = 4,777.1.

LC and two direct effects, led to splitting up the large first class from the initial solution. The second group-level class in this model is similar to the second class in the model presented in Table 2.5. The first class from Table 2.5, however, is split up into two classes. These two classes are rather similar when compared with each other, as they are when compared with the class from the first model, but with a large difference in degree of task repetition reported by the team members.

The results from Table 2.11 clearly show the difficulty in capturing team differences using team-level classes, as the first and third class differ only with respect to the degree of task repetition. Given that the group-level classes in the initial model are affecting the indicators only indirectly through the lower-level LC, such a relatively small difference between teams may become obscured between other characteristics that the teams do have in common. That is, detecting these specific characteristics on the team level in a model without direct effects from the team-level latent variable also requires more classes on the lower level. Such an addition of LCs on either level is not warranted when inspecting the BIC values for these models, which are known to favor model parsimony. However, through the proposed BVR-group and BVR-pair this lack of a direct effect between the group-level LC and the repetitiveness variable could be detected, as well as the subsequent need for an additional class on the group level.

Maybe more important, because of the improved fit and the possibility to test assumptions, the model arrives at different substantive results. In this instance, the added group-level class causes a separation based primarily on task repetitiveness. Given that the interest lies in relating the classes to job satisfaction or turnover intent as an outcome, the results may differ between the original model as depicted in Table 2.5, and the better fitting model arrived at in Table 2.11. When, for example, task repetitiveness on the team level is detrimental to employee job satisfaction, it would have been hard to distinguish as an important factor in the model with two group-level classes. It would, however, be visible in the model with three group-level classes in which a comparison between the first and third group-level classes would identify repetitiveness as an important factor.

Using the residuals as additional guidance now results in a model with substantial better fit that would likely not have been found when relying only on the BIC

TABLE 2.9: Residuals for the three class, three group-level class model. Covariance between Variation and Diverse and direct effect from the group-level latent variable on Nonrepetitive. Bootstrap p-values between parentheses

	Nonrepetitive	Creative	Diverse	Capacities	Variation
Creative	0.073 (.720)				
Diverse	0.315 (.214)	0.054 (.362)			
Capacities	0.620 (.274)	0.170 (.536)	0.003 (.880)		
Variation	0.046 (.378)	0.114 (.154)	0.000 (.999)	0.053 (.546)	
BVR-group	1.041 (.046)	1.185 (.012)	0.843 (.316)	1.150 (.054)	0.931 (.290)
BVR-pair	0.138 (.214)	0.589 (.020)	0.051 (.496)	0.326 (.118)	0.092 (.454)

Note: Bayesian information criterion = 4,768.9.

TABLE 2.10: Residuals for the three class, three group-level class model. Covariance between Variation and Diverse and direct effect from the group-level latent variable on Nonrepetitive and Creative. Bootstrap p-values between parentheses

	Nonrepetitive	Creative	Diverse	Capacities	Variation
Creative	0.001 (.950)				
Diverse	0.530 (.108)	0.085 (.192)			
Capacities	0.837 (.186)	0.005 (.858)	0.090 (.454)		
Variation	0.003 (.890)	0.048 (.238)	0.000 (.999)	0.023 (.716)	
BVR-group	0.771 (.260)	0.739 (.452)	0.927 (.112)	1.083 (.136)	0.914 (.216)
BVR-pair	0.016 (.628)	0.011 (.696)	0.202 (.174)	0.280 (.150)	0.014 (.728)

Note: Bayesian information criterion = 4,775.3.

or comparable criteria. Both the proposed BVR-group and BVR-pair in combination with the BVR, allow the detection of the initial assumption violations, and they identify not only which part of the model but also which specific parameters may prove problematic. Misfit can be pinpointed and tested, allowing far more informed and directed model adjustments, which may lead to different, more thoroughly tested, substantive results.

It must be pointed out that in the application, the residuals were used as guidance for illustration. However, comparable with many residual measures as well as modification indices, the measures are by no means tied to a certain solution and indicate only badness of fit and assumption violations. That is to say, model adjustments should be theoretically driven, and blind adjustments to the model with the mere goal of improving the fit should be discouraged as a poor research practice that may, for example, lead to capitalization on chance (e.g. Kaplan, 1990; MacCallum, Roznowski, & Necowitz, 1992).

2.5 Simulation

As a proof of concept, a small simulation study is presented in this section. The final model from the application is used as the population model in the two conditions presented, which contains three classes on both levels, a direct effect on the creative and nonrepetitive variables, as well as a covariance between the diverse and variation variables. The exact logit parameters for this model can be found in

TABLE 2.11: Profile for the three class, three group-level class model. Covariance between Variation and Diverse and direct effect from the group-level latent variable on Nonrepetitive and Creative. Bootstrap p-values between parentheses

	Group class 1 Repetitive	Group-class 2 Defined	Group-class 3 Nonrepetitive	Class 1 Diverse	Class 2 Structure	Class 3 Creative	Overall
Nonrepetitive	.301	.316	.613	.554	.130	.233	.400
Creative	.660	.348	.674	.731	.077	.844	.557
Diverse	.822	.521	.754	.953	.209	.526	.698
Capacity	.753	.590	.707	.851	.444	.342	.683
Variation	.786	.506	.704	.962	.263	.000[a]	.665
Class 1	.784	.382	.678				
Class 2	.122	.529	.195				
Class 3	.095	.090	.127				
Prevalence	.352	.345	.302	.613	.284	.103	

[a] Boundary solution.

Appendix C. LatentGOLD 5.0 is used to generate 500 replicate data sets, which are subsequently analyzed using the correct model, and a misspecified model that does not contain the covariance and direct effects, and has only two group-level classes. Identical to the application, bootstrapped p-values are obtained on the basis of 500 iterations.

For the misspecified model, the expectation is that the BVR-group detects the absence of the two direct effects. Table 2.12 shows that the power to detect one of these misspecifications indeed turns out to be high (.788). The second direct effect is not detected. However, the logit parameters of the direct effect on the creative variable are minute (effect coded –0.186 and –0.005). In addition, the type I error for the three other variables does not differ significantly from the alpha level. The power to detect the missing class through the BVR-pair is equally high. Tables 2.5 and 2.11 show that the major change between the two- and three-class models is a separation of classes solely on the basis of the nonrepetitive variable. When the classes are not separated, BVR-pair detects the residual dependence between respondents on this particular variable. That the values for the variation variable are higher than the nominal alpha levels can be explained by the fact that the logit parameter in the population model is extremely high. As a result, in an attempt to explain maximum group-level dependence, the model underestimates the dependence resulting from the variation variable to be able to explain the group-level dependence that results from the nonrepetitive and creative variables.

Table 2.13 shows the proportion of rejections under the correctly specified model, that is, the type I error. Satisfactory error rates should be close to the nominal .05 level. This is the case in most instances, but BVR-pair for the diverse variable, as well as BVR-group for the creative variable, differs significantly from .05. Whether this is the result of the additional direct effect and covariance for these variables and is a systematic issue requires further study and a more extensive simulation study. The absolute differences appear to be small, however.

To summarize, although the simulation presented here is necessarily limited in

TABLE 2.12: Misspecified model: Proportion of replications with bootstrap p-values <.05 (Power)

	Nonrepetitive	Creative	Diverse	Variation	Capacities
BVR-group	0.788	0.064	0.050	0.376	0.052
BVR-pair	0.872	0.048	0.032	0.378	0.042

scope, the power of the introduced measures is high, and the type I error rates are close to their nominal levels. This simulation therefore demonstrates that our measures' performance is satisfactory in the case of the application discussed, and it provides proof of concept from which future investigations may depart.

2.6 Discussion

Several problems occur when using only global fit statistics or information criteria for model selection in multilevel LC analysis. Because of the lack of local fit statistics, potential model misfit may go unnoticed, and there is no information available regarding how a model might be adjusted and improved. Therefore two new local fit statistics, BVR-group and BVR-pair are proposed, which test individual areas of the model and as such help in determining which areas of the model are problematic and how a model can best be improved. In conjunction with the standard BVR, they also allow the two local independence assumptions central to multilevel LC models to be inspected and tested. Computation of both the BVR-group and BVR-pair is already implemented in the user-friendly LatentGOLD 5.0 software package.

By using the BVR-group and BVR-pair as additional guidance to test and improve a multilevel LC model, it is shown that they enhance the ease with which fruitful model adjustments can be found. The model obtained by relying on the two residuals has better global fit and is known to better adhere to the local independence assumptions. The usefulness of the residuals is further emphasized by the change in substantive results between the initial model selected through the BIC and the latter model as improved through the use of the proposed statistics. That is, the misfit that is detected in this instance is not a mere misspecification against which the model is robust but actually distorts model-based conclusions.

That these model improvements can be found using a stepwise approach, and that such an approach may lead to finding relations and effects that would otherwise go unnoticed, does not, however, mean that these improvements lead to the true population model. It should be noted that there is a substantial risk for capitalization on chance, and in practice, such an approach should be used in conjunction with a form of replication such as cross-validation. That is, the residuals merely indicate local misfit and do not point to a given solution for that particular misfit.

Nonetheless, in this case important sources of misfit that affect the results have been picked up by the two residuals. Still, this chapter serves as an introduction, and a more in-depth simulation study is lacking. Such a future study would not focus as

TABLE 2.13: Correct model: Proportion of replications with bootstrap p-values <.05 (Type I error)

	Nonrepetitive	Creative	Diverse	Variation	Capacities
BVR-group	0.034	0.028	0.044	0.034	0.040
BVR-pair	0.046	0.052	0.076	0.046	0.048

much on the type I error rates, as the process of p-value bootstrapping is identical to that of the BVR for which it has been extensively tested (Oberski, Van Kollenburg, & Vermunt, 2013). Rather, it would focus on the consistency with which misspecification is detected under different circumstances and in more complex models, such as models incorporating covariates.

An additional extension that does require future work is to develop a similar residual for LC models for longitudinal data, in which dependencies can be assumed to take on the form of autocorrelation structures. Furthermore, the use of the BVR-group and BVR-pair may be studied for different methods and models that could also benefit from these statistics (e.g. see Varriale & Vermunt, 2012). The residuals are developed with the aim of testing the local fit of multilevel LC models, but they can be applied to all cases in which categorical multilevel data are used. The observed frequencies would be identical when applying the residuals to other methods, and only the expected frequencies would need to be obtained from the alternative approach.

Chapter 3

Power and Type I Error of Local Fit Statistics in Multilevel Latent Class Analysis

Abstract

In the social and behavioral sciences, variables are often categorical and people are often nested in groups. Models for such data, such as multilevel logistic regression or the multilevel latent class model, should account for not only the categorical nature of the variables, but also the nested structure of the persons. To assess whether the model accomplishes this goal adequately, local fit measures for multilevel categorical data were introduced in Chapter 2 (published as Nagelkerke, Oberski, & Vermunt, 2016). The BVR-group evaluates the variable-group fit, while the BVR-pair evaluates the person-person fit within groups. In this chapter, we evaluate the performance of these two measures for the multilevel latent class model (Vermunt, 2003). An extensive simulation study indicates that whenever multilevel latent class modeling itself is viable, type I error is controlled and power adequate for both fit statistics. Thus, the BVR-group and BVR-pair are useful measures to locate important sources of misfit in multilevel latent class analysis.

This chapter is published as: Nagelkerke, E., Oberski, D. L. & Vermunt, J.K. (2017). Power and Type I Error of Local Fit Statistics in Multilevel Latent Class Analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), pp.216-229.

3.1 Introduction

Latent class (LC) models can be used to search for classes of systematically similar respondents by considering their responses to a number of discrete indicator items. Analogous to many statistical methods this model assumes the observations that are classified to be independent. However, dependence often does occur when respondents are observed in naturally occurring groups, leading to a violation of the assumption. When ignored, this dependence will bias the results (Park & Yu, 2016). The multilevel extension to the LC model provides a solution for such cases of nested categorical data (Vermunt, 2003) by taking the grouping into account. Additionally, and maybe more important, it does not only solve the statistical problem of dependent observations, but it substantively allows observed groups to be classified based on their members (Vermunt, 2003, 2008) providing a simultaneous classification of individuals and groups.

The resulting classification of respondents and groups could be used as a predictor in subsequent analyses (e.g. Roosma, Van Oorschot, & Gelissen, 2016), or covariates can be added to the model to try and substantively explain the classes after an exploratory or confirmatory classification (e.g. Fagginger Auer, Hickendorff, Van Putten, Béguin, & Heiser, 2016; Tomczyk, Hanewinkel, & Isensee, 2015). Regardless of the approach, in both these cases the quality of the classification has a direct influence on the quality of the eventual outcomes of interest, and the fit of the measurement model should be carefully considered before continuing with further analyses.

Central to the model fit in multilevel LC analysis are two assumptions of conditional independence given the latent variables. On the lower level the assumption is that all dependence between items is captured by the latent variable, thus assuming conditional independence of the indicators given the LC variable. This assumption is identical to that of a regular LC model. On the higher level a similar assumption is made, where the observed group members are assumed conditionally independent given the higher-level latent variable.

In such relatively complex models, with assumptions on two levels and the distinct substantive goals of reproducing the overall and within-group responses, local fit statistics are of increased importance. Traditionally, global fit indexes such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are used to examine whether all of the assumptions hold, relative to some measure of model complexity. However, this approach has the disadvantage that misspecifications that are small relative to the model complexity may in fact still be harmful to subsequent analyses of interest (Oberski, 2014). Furthermore, their use generally limits itself to the comparison of estimated models in practice. With a nigh infinite number of model specifications, the selected, best fitting model out of the estimated alternatives might very well contain misspecifications and assumption violations. For this reason, the global fit measures can best be supplemented with measures of

local fit that examine the strength of evidence against individual model assumptions.

To examine local fit in models for multilevel categorical data, in Chapter 2 the BVR-group and BVR-pair measures are proposed. Both are in line with the bivariate residual (BVR) proposed by Vermunt & Magidson (2013) that measures how well the item-item dependence is captured by a single-level LC model. The two multilevel fit measures are comparable, but test how well the model captures the group-item dependence and person-person dependence related to the higher level of the model. All three, the BVR-group, BVR-pair, and BVR, take the form of a Pearson residual, but despite this resemblance they do not follow an asymptotic chi-square distribution. P-values can nonetheless be obtained relatively easily by means of a parametric bootstrap (Oberski, Van Kollenburg, & Vermunt, 2013).

The two higher-level residuals respectively aim to detect misfit related to the conditional independence assumption and substantively correct reproduction of the data. The BVR-group signals residual dependence between observed group membership and indicator items. When such residual dependence exists it is an indication of the model not fitting one or more of the groups correctly, implying that the model does not fully capture the between-group differences. The BVR-pair signals residual dependence between persons that are members of the same group. This residual dependence is also indicative of the model not correctly capturing the nested structure of the data, but here the focus is on the within-group similarities of the group members.

In Chapter 2 only a limited simulation study is provided, however, and little is currently known about the properties of the two statistics. With an extensive simulation study we here aim to more thoroughly investigate the power and type I error of the bootstrapped BVR-pair and BVR-group. Of primary interest is whether and under what conditions the two statistics have enough power to detect several types of misspecification of the multilevel LC model. The misspecifications of the model that are considered are closely related to the two assumptions of conditional independence added by the multilevel extension to the LC model; that is, the assumption that the members of an observed cluster in the data are independent conditional on the higher-level latent variable to achieve a group-level classification, and the assumption that observed group membership and the individual responses are conditionally independent to correctly reproduce the observed responses within the observed groups (Vermunt, 2003).

It should be noted that the context of the study is confined to multilevel LC analysis for which the statistics are originally developed, but that they can be obtained for any method that models nested categorical data, such as multilevel IRT. The two residuals namely aim to test for the correct modeling of within-group similarities and between-group differences by contrasting the observed and expected frequencies. Whether these expected frequencies are obtained from a multilevel LC model or an alternative method does not impact the way in which the eventual values are obtained.

The remainder of this chapter is structured as follows. In the following section the multilevel LC model is briefly introduced. Next the BVR-group and BVR-pair statistics are described, after which the design of the simulation study, including the bootstrap procedure are discussed. The results of the simulation study and the conclusions that can be drawn in terms of type I error and power are presented in the final two sections.[1]

3.2 The Multilevel Latent Class Model

The multilevel LC model is described using two equations. Both strongly resemble the expression of a regular LC model which classifies individuals based on the probabilities of their responses. The equation for the lower level of a multilevel model does exactly the same, but to take into account the nested structure of the data the response probabilities are made conditional on the group-class membership. To classify the groups and obtain this group-class membership, the higher-level equation describes the marginal probabilities of the combined response patterns of the group members of observed groups; that is, it describes the vector of response patterns that is obtained by combining all members of a group (Vermunt, 2003, 2008).

Let the lower-level latent variable be denoted as η_{ij}, classifying units in C latent classes, with one class referred to as c. The higher-level latent variable is denoted ζ_j, with G group-level latent classes, one of which is denoted g. Here the response of individual i in group j to item k is denoted y_{ijk}, with a total of J groups, all having n_j members summing to N, and K items having R_k categories. The vector of responses of individual i in group j to all K items is denoted $\mathbf{y}_{ij}$, with $\mathbf{r}$ referring to a particular answer pattern and r_k referring to one particular response to item k. Assuming conditional independence the lower level of the model is expressed as:

$$Pr(\mathbf{y}_{ij} = \mathbf{r}|\zeta_j = g) = \sum_{c=1}^{C} Pr(\eta_{ij} = c|\zeta_j = g) \prod_{k=1}^{K} Pr(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g). \quad (3.1)$$

When the conditioning on the group-level latent variable is removed, Equation 3.1 is identical to that of a regular LC model. Without this conditioning the probability of observing a certain pattern of responses $\mathbf{r}$ is the sum over the unconditional probability of class membership multiplied by the product of all conditional probabilities of observing the separate responses r_k. In turn conditioning all these terms on the group-level classes $(\zeta_j = g)$ allows the classification of groups.

Given the lower-level expression, the higher level now describes the classification of groups based on their members. Here the vector of all response patterns of units within group j is denoted as $\mathbf{y}_j$, with $\mathbf{s}$ denoting a particular combination of

[1] Appendices and additional resources can be found online at the Open Science Framework, at the permanent URL: osf.io/23mp2.

response patterns. The conditional independence assumption on this level relates to the units within groups, where not the responses to single items, but the entire response patterns of group members are assumed independent (Vermunt, 2003). The upper level can then be expressed as:

$$Pr(\mathbf{y}_j = \mathbf{s}) = \sum_{g=1}^{G} Pr(\zeta_j = g) \prod_{i=1}^{n_j} Pr(\mathbf{y}_{ij} = \mathbf{r}|\zeta_j = g). \tag{3.2}$$

This second equation likewise resembles that of a regular LC model, but now the full vector $\mathbf{y}_j$ of individual response patterns $\mathbf{s}$ in group j is described as a combination of the size, or prevalence, of group-level LC g, and the conditional probabilities of observing the combination of individual answer patterns $\mathbf{r}$.

Equations 3.1 and 3.2 describe the most general form of the multilevel LC model, in which both the class sizes and the response probabilities are allowed to vary across group-level LCs. This general form is hardly ever used, because of the difficulty interpreting group clusters with a completely different lower-level structure. There are two common ways to constrain the model. Setting $Pr(\eta_{ij} = c|\zeta_j = g) = Pr(\eta_{ij} = c)$ fixes the class membership on the lower level to be independent of that on the higher level, but allows the response probabilities to be estimated freely. The second and most common constraint $Pr(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) = Pr(y_{ijk} = r_k|\eta_{ij} = g)$ inversely fixes the response probabilities on the lower level to be independent of the higher-level class membership, but allows the class sizes to be estimated freely. The latter constraint leads to the model that simultaneously classifies respondents and the groups in which they are nested (Lukočiené, Varriale, & Vermunt, 2010).

3.3 Multilevel Local Fit Statistics

The idea behind the two fit statistics for the higher level of the multilevel LC model is relatively straightforward. Given that LC analysis is concerned with categorical indicators, and both the substantive goal of the model as well as the assumptions it makes can be reduced to adhering to a conditional independence assumption, a test comparable to a chi-square test is an intuitive solution. Both statistics can then compare the dependencies captured by the model to the dependencies present in the data. Because the asymptotic distribution is unknown, in the following the type I error and power are considered for the bootstrap of the measures.

3.3.1 BVR-group Residual

The BVR-group is concerned with the average model fit across groups, and quantifies the covariance between observed groups and items that is not captured by the model. When such residual covariance exists the observed group membership still affects the response probabilities of group members, implying that between-group

differences are not fully captured by the group-level latent variable; that is, the BVR-group tests whether the observed response frequencies within the observed groups are adequately reproduced by the model of interest.

The expectation under a well fitting model is, of course, that the expected and observed response frequencies are close to identical. For the within-group frequencies this implies that, given the model, the indicator variables should be conditionally independent of observed group membership. To test this, a Pearson residual can be obtained by cross-tabulating the observed and expected frequencies within all groups. This residual is then indicative of all the uncaptured variation caused by observed group membership.

The expected frequency, denoted as m_{jr}, can be obtained from the model as the individual probability of giving a certain response $Pr(y_{ijk} = r_k)$ and summing this probability over the group members:

$$Pr(y_{ijk} = r_k) = \sum_{g=1}^{G} Pr(y_{ijk} = r_k|\zeta_j = g)Pr(\zeta_j = g|\mathbf{y}_j = \mathbf{s}), \tag{3.3}$$

with

$$Pr(y_{ijk} = r_k|\zeta_j = g) = \sum_{c=1}^{C} Pr(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g)Pr(\eta_{ij} = c|\zeta_j = g). \tag{3.4}$$

Then,

$$m_{jr} = \sum_{i=1}^{n_j} Pr(y_{ijk} = r_k). \tag{3.5}$$

These equations can be simplified in a model without covariates to multiplying the probability of a response with the number of group members, rather than the more general sum over n_j in Equation 3.5. The observed response frequencies, denoted n_{jr}, are a simple count of the responses given by the group members. The BVR-group then equals:

$$BVR_{group.k} = \frac{1}{(R_k - 1)(J - 1)} \sum_{j=1}^{J} \sum_{r=1}^{R_k} \frac{(n_{jr} - m_{jr})^2}{m_{jr}}. \tag{3.6}$$

As shown in Equation 3.6 a separate residual is computed for each group and each response category, all of which are subsequently summed over the J groups, and R_k categories. Additionally, the resulting statistic is divided by $(R_k - 1)(J - 1)$, which is the number of non-redundant parameters in the cross-table, standardizing the BVR-group so it is not affected by the number of groups in the data and the number of categories of the variable.

Because the focus is mainly on item specific misfit, the BVR-group is here obtained per item. However, by removing the sum over J groups the statistic can be obtained per group to inspect whether misfit originates from the model not fitting

specific groups. Moreover, by not summing over the R_k categories it can be obtained per response category, which could be useful when extreme responses are a plausible cause of misfit.

3.3.2 BVR-pair Residual

On the higher level of a multilevel LC model the assumption is made that given the group-level latent variable the response patterns of nested units are conditionally independent. That is, the full response patterns $\mathbf{r}$ of all n_j group members in Equation 3.2 are assumed to be conditionally independent. The BVR-pair tests for violations of this assumption. When there is residual dependence among the members of observed groups the within-group similarity is not correctly reproduced by the group-level latent variable. In other words, the nested structure of the data is not fully captured by the model.

Because the assumption on this level does not relate to the items, but to the units, the group members need to be related to one another. This is done by creating all possible pairs of units within an observed group to obtain the pairwise response frequencies. When the assumption that all dependence between the units is captured by the model holds, the expected and observed frequencies would again be in agreement. Here, by considering the pairwise frequencies, this would be indicative of the response of unit i and i', rather than item k and k', being locally independent.

The expected frequency of a pair of responses is obtained using the joint probability of unit i giving response r, and unit i' giving response r' to item k:

$$Pr(y_{ijk} = r_k, y_{i'jk} = r'_k) = \sum_{g=1}^{G} Pr(y_{ijk} = r_k|\zeta_j = g)Pr(y_{i'jk} = r'_k|\zeta_j = g)Pr(\zeta_j = g|\mathbf{y}_j = \mathbf{s}). \quad (3.7)$$

After which the expected frequency $m_{krr'}$ can be obtained by multiplying with the number of possible pairs within the group:

$$m_{krr'} = \sum_{j=1}^{J}(n_j(n_j - 1)/2)Pr(y_{ijk} = r_k, y_{i'jk} = r'_k). \quad (3.8)$$

Essentially, the probabilities of all possible combinations of the discrete responses to a single item are obtained per group and multiplied by the number of possible pairs of members in a group.

Obtaining the observed frequency ($n_{krr'}$) can be thought of as creating a cross-table for each pair. This table would identify the combined response of unit i and i' to item k, since only one cell would have a value of one. Subsequently summing these tables over all pairs results in the pairwise frequency for that particular item (for an illustration, see Chapter 2 - Table 2.1).

Important to note here is that in a multilevel LC model the ordering of the responses does not matter for the probability of a pair. For example, two units responding to a dichotomous item forming a yes-no pair, have a probability that is identical to a no-yes pair. Yet, in practice the observed frequencies for such pairs will almost always differ depending on how the data set is ordered. Therefore, patterns with the same, but differently ordered responses are summed when obtaining the BVR-pair statistic:

$$BVR_{pair} = \frac{J}{N}\frac{1}{R_k(R_k-1)/2}\left[\sum_{r}^{R_k}\sum_{r'>r}^{R'_k}\frac{((n_{krr'}+n_{kr'r})-(m_{krr'}+m_{kr'r}))^2}{m_{krr'}+m_{kr'r}} + \sum_{r}\frac{(n_{krr}-m_{krr})^2}{m_{krr}}\right]. \quad (3.9)$$

To arrive at the BVR-pair, the raw residual is divided by the number of non-redundant parameters in the table. Given the symmetry on the off-diagonals this is a division by $R_k(R_k-1)/2$. Additionally, because the theoretical maximum value increases as a triangular sequence with n_j, the statistic is divided by the average group size, simply to reduce the resulting values.

Because of this triangular increase one more problem needs to be solved when the groups are of different sizes, as it causes units in larger groups to be in far more pairs than those in smaller groups. As a result, the observed and expected marginal frequencies can differ, where $(n_{krr'}+n_{kr'r}) \neq (m_{krr'}+m_{kr'r})$. Such a difference affects the values of the BVR-pair, whilst not being indicative of residual dependence. To avoid the influence of these marginal differences on the BVR-pair, iterative proportional fitting is used to update the table with expected pairwise frequencies so that it retains its cross-product ratios (Bishop, Fienberg, & Holland, 1975), but has the observed marginal frequencies (again, for an illustration see Chapter 2 - Table 2.3).

Note that the computational complexity of obtaining the BVR-pair is primarily determined by the number of items, possible responses, and group-level classes. For the BVR-group this would be the number of possible responses and groups. The sample size in terms of the number of observations and groups only affects frequency counts and thus adds little time to the required computations. However, because the residuals do not follow a known asymptotic distribution, a bootstrap is required to obtain p-values. This, of course, does increase the computational times, and may make obtaining the computationally intensive for truly big data sets. For $N = 62,500$ the average time for 250 bootstraps in this study was approximately 9 minutes on a 4 x 3.30 Ghz processor.

3.4 Simulation Design

The misfit that the two residual statistics aim to capture are the model not fitting observed groups, causing residual conditional dependence between group membership and indicator items, and the model not capturing all within-group dependence between units, causing a residual dependence between pairs of observations. These types of misift can be remedied in the multilevel LC model by either allowing a direct effect from the group-level LC variable on one or more of the indicators, or adding additional group-level LCs. To test the power of detecting such misfit this logic is reversed, whereby a population model is assumed containing for instance a direct effect and analyzing these data with a misfitting model excluding that particular parameter. To investigate the power and type I error of the two residuals a Monte Carlo simulation is used to evaluate a range of different models, with differing types of misspecification.

3.4.1 Variables and Factors

The power in LC models themselves is primarily dependent on two mutually influencing factors, namely the amount of information and entropy, or class separation. The former is what affects the power of any statistical test, and depends on commonly studied factors such as the sample size, the size of observed groups and the number of observed items. In LC analysis an important additional aspect is how distinctly different the LCs are, which is affected by the number of classes and the effect sizes of the parameters.
The factors that are varied and affect the structure of the sample are:

- Number of observed groups: 50, 100 or 250 groups.
- Number of observed group members: 10, 50, or 250 group members.
- Number of indicator items: 6 or 10 items.

The factors that are varied and can be thought of as model specific are:

- Number of lower-level classes: 2 or 3 classes.
- Number of higher-level classes: 2 or 3 classes.
- Loglinear effect from the lower-level latent variable to the indicator: 0.424 or 0.693 (conditional response probabilities).
- Loglinear effect from the higher-level latent variable on one or two indicators: 0.000 (no direct effect), 0.201 or 0.511.
- Loglinear effect from the higher-level latent variable on the lower-level latent variable: See Table 3.1.

The loglinear parameters are effect coded, leading to conditional probabilities of 0.7 or 0.8 in one lower-level class, and the complement of 0.3 or 0.2 in the other. In conditions with three classes, half of the items in the middle class have a conditional probability of 0.7 or 0.8, and the complement for the other half of the items. For

TABLE 3.1: Logit parameters for the higher level: Effects of the first group-level class on the two or three lower-level classes [a]

	2 lower-level classes			3 lower-level classes		
	Logit 1	Logit 2	Logit 3	Logit 1	Logit 2	Logit 3
2 group class (W)	0.424	-0.424	-	0.196	0.014	-0.209
3 group class (W)	0.424	0.000	-0.424	-0.514	1.027	-0.514
2 group class (S)	0.693	-0.693	-	0.928	0.341	-1.269
3 group class (S)	0.693	0.000	-0.693	-0.693	1.386	-0.693

a. For examples of the resulting conditional probabilities see Appendix D. For all models see the Open Science Framework: osf.io/23mp2.

examples of the conditional probabilities in the different population models see Appendix D. All intercept values are kept at zero, which implies equal class sizes. Of course, by crossing the number of groups and their members different sample sizes are obtained, namely 500, 1000, 2500, 5000, 12,500, 25,000, and 62,500.

The power to detect misfit is considered for eight types of misspecification, as well as for the correctly specified models to estimate the type I error.
The misspecifications considered are:

- A missing class on the lower level
- A missing class on the higher level
- A missing direct effect (weak and strong)
- A missing direct effect when there are two direct effects (weak and strong)
- A missing class on the higher level and a missing direct effect

Design of Experiments

It needs to be taken into account that the model itself is relatively complex, and that the residuals require a parametric bootstrap. This leads to many model re-estimations since the bootstrap needs to be performed for each Monte Carlo replication. To reduce the computational intensity and keep the study feasible, a smaller design than full factorial was chosen, whereby the higher order interactions between the variables are deliberately left confounded (see e.g. Lundstedt et al., 1998). The idea is identical to a fractional factorial design, or I^{k-p} design, but because the variables of interest have different numbers of levels the setup does not result in a true fraction of the full factorial. Using SAS JMP (see e.g. Montgomery, 2012) a design consisting of 422 conditions was generated that has no aliasing for the main effects, nor for the second and third order interactions in the full set of conditions. This way, only one fifth of the computations are needed. The compromise is that higher order interactions cannot be estimated, although generally four variable interaction effects and up are of limited practical value. It must be noted that these are interactions on the variable level, which means that the limitations occur on the factor level, where certain combinations are not taken into account. For example, all low N conditions have an observed group size of 10.

3.4.2 Monte Carlo and Bootstrap

The Monte Carlo simulation is conducted using a combination of R (R Development Core Team, 2015) and LatentGOLD 5.0 (Vermunt & Magidson, 2015), whereby R is used to generate syntax and post-process the results. Based on the desired population model, LatentGOLD is used to generate a data set, which is subsequently analyzed with either a correctly or misspecified estimation model. To obtain the p-value for the BVR-pair and BVR-group statistics a bootstrap is conducted using the maximum likelihood values that follow from the estimation.

The bootstrap data is obtained by sampling group-level LC membership based on the class prevalences, class membership conditional on the sampled group-level LC membership, and finally the responses conditional on both the sampled memberships. The p-value for the BVR statistics are then obtained by computing the proportion of bootstrap samples in which the residuals are larger than in the original model. This process of generating data, analyzing the data, and performing the bootstrap is repeated for the desired number of Monte Carlo replications. The proportion of significant p-values of the total number of replications then is indicative of the power. For the null-models both the number of bootstrap samples and Monte Carlo replications are set to 250. For the misspecified models both are set to 500 for the large majority of models, with the exception of several conditions with a very large N and weak class separation that are computationally extremely intensive.

3.5 Results

First the results for the null-models will be discussed, since estimations of power cannot be interpreted when the nominal alpha levels are incorrect.

3.5.1 Type I Error

Table 3.2 depicts the average proportion of significant BVR values at the $\alpha = .05$ level for the first indicator variable when a direct effect is present, and the third when there is not, where the mean is computed over all conditions that satisfy a particular factor. Note that this reverses the interpretation of the numbers in the table, where all values lower than .05 are too liberal, since there are too few significant values indicating misfit. The reason for depicting the third indicator is that, when present, direct effects from the group-level latent variable on an indicator are on indicators one and/or two. L here refers to the number of conditions that the average is based on, because not all factors occur equally often due to the study design.

Overall, the BVR-group and BVR-pair statistics are very close to the nominal alpha level, regardless of the condition that the mean is computed over. The BVR-group, however, is slightly too liberal, especially in the conditions with a smaller N. This may for a large part be due to the statistic taking the form of a chi-square (χ^2)

TABLE 3.2: Type I Error: BVR-group and BVR-pair mean and standard deviation of the proportion of significant bootstraps per main factor for an indicator item with and without direct effects

	With a direct effect on the item					Without direct effect on the item				
	L	BVR-group	SD	BVR-pair	SD	L	BVR-group	SD	BVR-pair	SD
Classes = 2	159	.040	.020	.047	.015	216	.049	.014	.050	.013
Classes = 3	156	.048	.015	.050	.013	206	.048	.016	.051	.014
Group Cl. = 2	158	.047	.015	.051	.012	208	.050	.014	.051	.014
Group Cl. = 3	157	.041	.020	.047	.015	214	.048	.016	.050	.014
Items = 6	160	.044	.018	.048	.014	215	.048	.016	.049	.014
Items = 10	155	.044	.018	.049	.015	207	.049	.014	.051	.013
Groups = 50	103	.044	.018	.049	.015	137	.049	.016	.050	.013
Groups = 100	103	.042	.017	.046	.013	142	.048	.015	.050	.013
Groups = 250	109	.046	.020	.051	.015	143	.048	.014	.051	.015
Members = 10	106	.034	.021	.046	.016	143	.044	.016	.051	.012
Members = 50	103	.048	.015	.050	.013	139	.050	.013	.051	.014
Members = 250	106	.050	.013	.050	.014	140	.052	.014	.049	.014
N = 500	35	.036	.020	.048	.018	48	.047	.017	.049	.011
N = 1000	33	.029	.017	.041	.012	47	.042	.016	.051	.012
N = 2500	73	.042	.021	.050	.014	95	.047	.015	.051	.014
N = 5000	34	.045	.014	.048	.012	46	.050	.013	.050	.013
N = 12500	67	.049	.015	.049	.013	88	.050	.014	.051	.015
N = 25000	36	.051	.012	.049	.012	49	.053	.013	.050	.013
N = 62500	37	.051	.014	.053	.015	49	.051	.013	.049	.017
Class Sep. = Low	155	.045	.019	.050	.015	205	.049	.015	.050	.013
Class Sep. = High	160	.043	.017	.047	.012	217	.048	.014	.051	.014
Group Sep. = Low	156	.043	.019	.049	.015	209	.048	.015	.051	.014
Group Sep. = High	159	.045	.017	.048	.013	213	.049	.014	.050	.013
Overall	315	.044	.018	.049	.014	422	.049	.015	.050	.014

test, which becomes more conservative as sparseness increases, also when a parametric bootstrap is used (Langeheine, Pannekoek, & Van de Pol, 1996; Von Davier, 1997). That is, the chi-square test is too conservative in that the null-hypothesis that there is no misfit is not rejected, making the BVR-group too liberal. In these conditions the number of groups is set to 50 or 100 with only 10 members, leading to relatively sparse frequency tables. This is in line with the BVR-pair not showing any problems, as it is obtained on a $R \times R$, rather than a $R \times J$ table, in addition to the number of pairs being far larger than the number of observations.

The left hand side of Table 3.2 depicts the type I error for the first indicator item and only for conditions in which a direct effect on the indicator is present. A direct effect being present causes slightly more variation, but the overall results are still good in terms of the type I error rate. The most problematic cases are clearly those where little information per group is available, especially when there are many small groups. This can for example be seen from the conditions $N = 500$ and $N = 1000$, both of which have 10 observed cases per group. Again, this can largely be attributed to sparseness.

TABLE 3.3: Power to detect ignoring the nested structure: The last three columns indicate the power to reject fit for item one, at least one item and at least half of the items, respectively

Sample			Class Separation				BVR-group		
N	Groups	Group Size	Items	Lvl.1	Lvl.2	C	Item 1	Min.1	50%
500	50	10	10	L	L	3	0.062	0.476	0.002
500	50	10	6	H	L	3	0.114	0.464	0.024
500	50	10	6	L	H	3	0.654	0.992	0.830
500	50	10	10	H	L	2	0.780	1.000	0.906
1000	100	10	10	L	L	2	0.468	0.996	0.582
1000	100	10	6	L	H	2	0.958	1.000	1.000
1000	100	10	10	H	H	2	1.000	1.000	1.000
2500	50	50	10	H	H	3	1.000	1.000	1.000
2500	50	50	10	L	H	2	1.000	1.000	1.000
2500	250	10	10	L	H	3	0.998	1.000	1.000
2500	250	10	6	H	H	3	1.000	1.000	1.000
5000	100	50	10	L	L	3	0.276	0.914	0.098
5000	100	50	6	H	L	3	0.656	0.986	0.862
5000	100	50	6	H	L	2	1.000	1.000	1.000

Inspecting the BVR that tests the local independence between items (not reported) on the lower level of the model, does not indicate any problems with the model itself either. Where it may have been possible that strong group-level dependencies affect the fit or the fit statistics on the lower level, there is no evidence of this occurring.

3.5.2 Power to Detect Ignored Nesting

The most fundamental type of misspecification considered in the simulation study follows from specifying a model with too few classes on the group level when only two are present in population. This results in a model that ignores the nested structure of the data altogether. As can be expected, the parameter estimates and latent class solution in this situation are strongly biased, both in the parametric (Kaplan & Keller, 2011) and non-parametric (Park & Yu, 2016) multilevel LC model.

Table 3.3 depicts the power to detect the presence of an additional group-level class when only one is specified in the analysis, which is identical to specifying a regular LC model. The full conditions are presented here, because splitting on all factors would result in a largely empty table, whereas confounding any of the factors would not provide the full picture. All conditions with a larger sample size are omitted, as the power equals one.

Preferably the BVR-group and BVR-pair values should be significant for each separate indicator item when detecting a missing class. The dependence that is not captured by the model is namely affecting all of the indicators. However, it is not necessarily the case that none of the group-level dependence is modeled on the lower level, and the fit of some of the indicators may well be acceptable. Vice versa,

if only one or two of the indicator items were identified as not being reproduced correctly by the model, the conclusion of a missing class would probably not be drawn, and model improvements would focus primarily around these specific items. Therefore, what is reported in the table are three proportions for the BVR-group. Namely, the power when only looking at the first item, the proportion of Monte Carlo replications where at least one out of the K residuals is significant, and the proportion of replications where 50% or more of the BVR-group values are significant (so for 3 or 5 out of 6 or 10 indicator items). For conciseness, the BVR-pair values are included separately in Appendix Table E.1, because they show an identical pattern, albeit slightly less powerful.

The power of the BVR-group to detect that *something* is wrong when completely ignoring the nested structure of the data, while there are two group-level classes, is close to one in practically all situations. Judging from the second to last column of Table 3.3 only in two extreme situations the combined power over all indicator items drops below .90. In these two cases class separation on the group level is almost nonexistent as shown in Table D.2 and Table D.3, with an estimated entropy of 0.317 and 0.323 respectively. Combined with the small sample size and associated uncertainty about the classification the dependence can actually be modeled without a group-level class. The nested structure in these situations is only detected with a truly large sample (power equals one in the omitted conditions with $N \geq 12,500$).

However, when misfit on any one of the items is detected, it is not necessarily the case that misfit is found for all separate items. Generally more than half of the items will be reported as problematic, but two remarkable discrepancies are the first $N = 1000$ and $N = 5000$ conditions. At least one BVR-group value is significant for these conditions, but rarely more than half indicate misfit. Inspecting these two conditions further the average number of significant BVR-group values over all the Monte Carlo replications are 4.91 and 2.50 out of 10, so it is still likely that misfit in multiple indicators is detected for these two cases, although it may be too few to point to an unmodeled group-level class.

In practice, this means that the BVR-group will detect the nested structure in more typical situations where N is not too small and class separation at the group level not too low. Situations with small N and an extremely low class separation at the group level should already be cause for concern in the sense that there might not be a nested structure strong enough to model. In all other situations, at least one of the BVR-group values will generally be significant with an $N \geq 1000$. Given that this is a situation where one (identical to no) group-level class is modeled, there is no other way to address this dependence than adding group-level classes. The exact number of significant BVR-group values is less relevant in this respect, but will be returned to in the next section.

TABLE 3.4: Power to detect a missing group-level class: The last three columns respectively indicate the power to reject fit for item one, at least one item and at least half of the items

Sample			Class Separation				BVR-group			
N	Groups	Group Size	Items	Lvl.1	Lvl.2	C	Item 1	Min. 1	50%	
500	50	10	6	L	H	2	0.062	0.238	0.002	
500	50	10	6	H	H	3	0.568	0.998	0.962	
500	10	50	6	L	H	2	0.104	0.504	0.020	
500	10	50	6	H	H	3	0.476	0.972	0.956	
1000	100	10	6	H	L	2	0.044	0.224	0.004	
1000	100	10	10	L	H	2	0.046	0.434	0.000	
1000	100	10	6	L	L	3	0.220	0.806	0.120	
1000	100	10	10	H	H	3	0.560	1.000	1.000	
2500	250	10	6	L	L	2	0.040	0.242	0.004	
2500	250	10	10	H	L	2	0.052	0.414	0.000	
2500	250	10	6	H	H	2	0.110	0.476	0.016	
2500	250	10	6	L	L	3	0.440	0.984	0.540	*
2500	250	10	10	L	H	2	0.064	0.528	0.000	*
2500	250	10	6	H	L	2	0.036	0.248	0.004	*
2500	250	10	10	H	H	3	0.556	1.000	1.000	*
2500	50	50	6	H	L	2	0.202	0.728	0.124	
2500	50	50	6	L	H	3	0.530	1.000	1.000	
2500	50	50	10	H	H	3	0.542	1.000	1.000	
2500	50	50	10	L	L	3	0.554	1.000	0.998	
5000	100	50	6	L	L	2	0.106	0.498	0.018	
5000	100	50	10	H	L	2	0.326	0.958	0.274	
5000	100	50	6	H	L	3	0.496	1.000	1.000	
5000	100	50	6	H	H	2	0.982	1.000	1.000	

3.5.3 Power to Detect a Missing Group-level Class

A logical next step to consider is the situation in which too few, rather than no, group-level classes are specified. From Table 3.4 it is evident that the power to detect a third group-level population class as missing when two are specified is markedly lower. Inspecting the conditions more closely the power of the BVR-group is acceptable in conditions with larger separation between the classes. Note here that separation on the lower level also directly affects separation on the higher level, as can be seen by the conditional probabilities in the population models, illustrated by the group-level classes in Table D.4 and D.5. The stronger dependence between group membership and the responses of its members in turn leads to a higher residual dependence when not modeled correctly. For the BVR-pair results see Table E.2

In case the classes are not as strongly separated, more information is required to detect that the population may contain an additional class. However, this is not achieved by simply having a larger sample, but requires the sample size at either level to be sufficient. That is, enough information needs to be available on both the higher and lower level to detect residual dependence on the higher level. This is not too surprising given the model specification, whereby observed groups are

essentially classified based on the lower-level class membership of their members. Although a similar sample size recommendation for multilevel LC analysis is not readily available, the consistently high power in conditions where group size is 50 is in line with previous research on multilevel logistic regression (Moineddin, Matheson, & Glazier, 2007).

To further clarify the mutual effect between the number of groups and their size several additional conditions were considered. The marked $N = 2500$ conditions in Table 3.4 are identical to the conditions with an N of 1000. Comparing these four conditions to the lower N ones clearly shows that increasing the number of groups when they are very small barely increases the power to detect the correct nested structure, whilst the sample size more than doubles. The $N = 500$ conditions show that conversely increasing the group size for a small number of groups does not increase the power in a similar fashion. Whether this is due to too little power of the multilevel LC model to detect the true structure, or the power of the BVR-group to detect the failure of modeling the true structure is hard to disentangle and both might be occurring.

A final remark on Table 3.4 is that the BVR-group residual is generally more powerful with three, compared to two lower-level classes, even when the higher-level classes are further apart in terms of conditional response probabilities. This is a general trend, which can best be explained in terms of the population data. When the group members belong to a higher number of distinct classes, the classification of the groups is automatically more fine grained as well. That is, there is a more diverse composition of the group members in terms of the lower-level class that they belong to. This diversity will create a larger effect of observed group membership on the probability to give a certain response, and hence, failing to model the effect will create a larger residual. Related to this, note that the number of indicator items in the condition are not further discussed, because it causes no systematic differences in the power estimates.

The model here turns out to be quite good at redistributing the residual dependence. The practical implication of these findings is that for weakly defined classes or samples with small groups residual dependence is not picked up by one particular item, or a large majority of the items. Although this implies that groups should have around fifty members, it may not actually be extremely problematic in terms of model adjustments. The dependence is truly redistributed and generally ends up in one or two items that do show problems. When these items are addressed, for example by allowing a direct effect between the item and the group-level latent variable, it will not resolve the problem and other indicators will show residual dependence (for an example see the application in Chapter 2, Tables 2.7 and 2.8). This will either cause many BVR-group values to start indicating problems, or iteratively cause a few to show problems until an additional group-level class is the best solution in terms of parsimony. Of course, addressing problematic items blindly to merely reduce the residual dependence does lead to capitalization on chance, and

will most likely not result in finding the population model. Given the results a good exploratory approach would be to use global fit statistic or information criteria to determine the number of classes, attempting to resolve any residual dependence with theoretically sensible parameters, and if the dependence returns in other indicators to increase the number of classes.

3.5.4 Power to Detect Missing Effects

A second general type of misspecification concerns a missing direct effect from the group-level latent variable to one of the indicators. This model mimics the situation in which observed group membership is not conditionally independent from the indicators, and the univariate item distributions are not properly reproduced by the model. Here the ideal outcome is reversed from the detection of a missing class in terms of the residuals, where the BVR-group and BVR-pair should only detect misfit in the item to which the direct effect pertains.

In Table 3.5 the power of detecting a weak missing direct effect is presented; that is, an effect that causes a small residual dependence between observed group membership and the first indicator item. With a few exceptions, the BVR-pair has notably higher power to detect the misspecification. A quick summary of the results is that power increases with sample size and is generally higher for larger, rather than more, groups. The latter is also confirmed by inspecting several additional conditions with ten groups with fifty members, otherwise identical to the $N = 500$ conditions presented, which all have slightly higher, but still insufficient power. An extra set of conditions is also used for the effect of having more indicator items, which increases power slightly. However, having four additional indicators primarily increases lower-level class separation, which in turn only substantially affects group-level class separation when the group-level effects are strong. That is, it primarily increases power in already high-power conditions, and has a limited effect on low-power conditions.

For the other factors, the results are somewhat paradoxical. First, it seems that in small sample conditions a stronger separation of the classes generally leads to lower power. However, this is an artifact of the importance of the direct effect to separate the classes. When the effect is highly important for class separation (i.e. creates a large discrepancy between the entropy of the model and the population) it is picked up in conditions with weakly separated classes as it creates very large residual dependencies. Furthermore, in conditions with more classes the power is generally lower. The reverse at first seems more likely, as there is more information on the correct specification. However, more classes simply make it easier to model dependencies as there are a lot more parameters that can be used to compensate for the missing direct effect.

It should be noted that the direct effect here is an effect coded logit of 0.201, which creates only very minor changes in conditional probabilities. The power to detect a missing direct effect with a stronger effect of 0.511, presented in Table E.3,

TABLE 3.5: Power to detect the absence of a weak direct effect from the group-level latent variable on the first indicator variable

Sample			Separation					Group-level Entropy		Lower-level Entropy		Power	
N	Groups	N_j	Items	Lvl.1	Lvl.2	C	G	Pop.	Model	Pop.	Model	BVR-group	BVR-pair
500	50	10	6	L	L	2	2	0.646	0.511	0.543	0.540	0.340	0.550
500	50	10	10	L	H	3	3	0.873	0.868	0.649	0.647	0.042	0.088
500	50	10	10	H	H	3	2	0.960	0.953	0.817	0.814	0.108	0.194
500	50	10	6	H	H	3	3	0.935	0.927	0.755	0.748	0.040	0.104
500	50	10	10	H	H	2	3	0.590	0.570	0.943	0.943	0.228	0.478
1000	100	10	10	L	L	2	2	0.684	0.585	0.704	0.703	0.680	0.918
1000	100	10	6	L	L	3	3	0.575	0.564	0.416	0.412	0.030	0.076
1000	100	10	6	L	H	2	2	0.858	0.818	0.613	0.619	0.090	0.574
1000	100	10	10	H	H	2	3	0.588	0.570	0.943	0.943	0.358	0.736
1000	100	10	6	H	H	3	2	0.913	0.915	0.677	0.675	0.084	0.222
2500	250	10	10	L	H	2	3	0.536	0.496	0.716	0.718	0.732	0.980
2500	250	10	6	L	H	3	2	0.728	0.721	0.419	0.412	0.040	0.102
2500	250	10	6	H	L	2	3	0.357	0.314	0.826	0.827	0.922	0.970
2500	250	10	10	H	L	3	3	0.874	0.871	0.830	0.828	0.084	0.614
2500	250	10	10	H	L	2	2	0.721	0.665	0.940	0.940	0.938	1.000
2500	250	10	10	H	L	3	2	0.258	0.150	0.787	0.787	0.058	0.064
2500	50	50	10	L	L	3	3	0.996	0.995	0.602	0.596	0.286	0.490
2500	50	50	6	L	L	3	2	0.639	0.279	0.333	0.331	0.728	0.674
2500	50	50	6	L	L	2	3	0.757	0.665	0.535	0.540	0.900	1.000
2500	50	50	10	L	L	2	2	0.993	0.983	0.709	0.714	0.910	1.000
2500	50	50	6	H	L	2	2	0.995	0.991	0.834	0.839	0.582	1.000
2500	50	50	10	H	H	2	2	1.000	1.000	0.949	0.949	0.378	1.000
2500	50	50	6	H	L	3	3	0.999	0.999	0.712	0.707	0.166	0.776
2500	50	50	10	H	H	3	3	1.000	1.000	0.868	0.866	0.154	0.826
5000	50	100	10	L	H	2	3	0.946	0.927	0.727	0.731	0.854	1.000
5000	50	100	6	L	L	2	2	0.990	0.972	0.554	0.565	0.708	1.000
5000	50	100	10	L	L	3	2	0.719	0.373	0.481	0.479	0.998	0.946
5000	50	100	6	H	H	2	3	0.959	0.951	0.845	0.849	0.650	1.000
5000	50	100	10	H	H	3	2	1.000	1.000	0.819	0.817	0.462	0.998
5000	50	100	6	H	L	3	3	0.999	0.999	0.712	0.708	0.286	0.972
12500	250	50	6	L	H	2	2	0.999	0.998	0.626	0.640	0.176	1.000
12500	250	50	6	L	H	3	3	0.997	0.997	0.596	0.587	0.134	0.698
12500	250	50	6	L	H	3	2	0.999	0.999	0.571	0.564	0.720	0.980
12500	250	50	10	H	L	2	3	0.816	0.775	0.938	0.939	1.000	1.000
12500	250	50	6	H	L	3	2	0.669	0.438	0.625	0.625	0.802	0.102

quickly approaches one for all conditions with an $N \geq 1000$. Only the conditions with two group-level and three lower-level classes remain an exception, but this is due to class separation being very low. See for example Tables D.2 and D.3, where it is debatable whether there is a nested structure at all.

In Table 3.6 the results are averaged for the different sample sizes. The average power seems relatively low, but this is due to a few conditions resulting in a power close to zero to detect the weak effect that is missing (see also Tabel 3.5). The last four columns give some insight into how precise the residuals are able to identify the problematic variable, as they should preferably not identify other indicators as causing misfit. The BVR-group here does surprisingly well, especially when considering that a direct effect from the group-level latent variable to any of the indicators

TABLE 3.6: Average power to detect the absence of a direct effect from the group-level latent variable on the first indicator variable, by sample size

					Item 1				Item 2			
					BVR-group		BVR-pair		BVR-group		BVR-pair	
N	Groups	Size	L 0.2	L 0.5	Log 0.2	Log 0.5	Log 0.2	Log 0.5	Log 0.2	Log 0.5	Log 0.2	Log 0.5
500	50	10	5	7	0.152	0.355	0.283	0.475	0.042	0.050	0.049	0.063
1000	100	10	5	6	0.250	0.849	0.505	0.916	0.044	0.056	0.038	0.090
2500	250	10	6	6	0.462	0.692	0.622	0.736	0.043	0.069	0.051	0.147
2500	50	50	8	5	0.513	0.702	0.846	0.936	0.059	0.090	0.136	0.242
5000	100	50	6	5	0.660	0.942	0.986	0.948	0.066	0.152	0.073	0.389
12 500	250	50	5	6	0.566	0.842	0.756	0.986	0.116	0.282	0.077	0.523
12 500	250	50	5	5	0.974	0.814	1.000	0.996	0.052	0.276	0.158	0.852
25 000	100	250	6	6	0.950	1.000	1.000	0.980	0.075	0.608	0.195	0.671
62 500	250	250	6	7	0.990	1.000	1.000	1.000	0.223	0.780	0.410	0.941

affects the LC solution (see e.g. Table D.6). When the direct effect in the population is strong enough, excluding it from the model will affect the conditional probabilities for all items in both the lower- and higher-level classes. In such a case one group-level class, and thus the members of the observed groups that are classified into that class, will systematically resemble one another more, causing the residual to report uncaptured dependence. This can readily be seen from the BVR-pair value for a strong direct effect and large N. Here the power is large enough to identify the additional dependence that is created between members of the same group by excluding a direct effect from the model, as the BVR-pair residual has a power of close to one to identify both the first and second indicator as problematic.

Yet, this does not occur as persistently as expected. In most of these conditions the BVR-group does not identify the second item as causing misfit up to a certain point. As explained, there is true uncaptured dependence in all indicators due to a missing direct effect, so it can be expected that as the amount of information to identify that dependence increases, such as having $N = 62,500$, it is indeed detected. Also, it cannot be expected that these residuals then remain equal to the nominal alpha level. Nonetheless, even with a power to detect the direct effect on the first indicator item of 0.9, mistakingly identifying the second indicator as problematic only occurs in less than 30% of the replications.

Finally in Table 3.7 the average power of the more powerful BVR-pair is shown for conditions where one direct effect is missing, but two are present in the population. Comparing the power to that of the BVR-pair for Log(0.5) effects in Table 3.6 it is clearly harder to detect this misspecification. Similarly comparing Item 3 in Table 3.7 to Item 2 in Table 3.6 the false detection rates go up slightly, which is not surprising given the stronger dependencies throughout the data. In large sample studies it is even the case that the BVR-pair values almost always indicate significant misfit on more than half of the indicator items, which could lead to the conclusion that there are too few group-level classes. This may, however, not be extremely problematic as it is unlikely that adding a group-level class will be able to fully resolve the residual dependence problem, and misfit will still be indicated for the first item.

TABLE 3.7: Average power of the BVR-pair to detect the absence of a direct effect on Item 1 when two are present, by sample size. The remaining effect is log(0.511) on Item 2 in all conditions

			Log 0.2 Missing					Log 0.5 Missing				
N	Groups	Size	L	Item 1	Item 2	Item 3	50%	L	Item 1	Item 2	Item 3	50%
500	50	10	7	0.243	0.045	0.059	0.004	4	0.538	0.055	0.114	0.020
1000	100	10	5	0.391	0.050	0.071	0.015	6	0.626	0.043	0.115	0.012
2500	250	10	7	0.651	0.042	0.089	0.017	6	0.659	0.043	0.155	0.095
2500	50	50	6	0.691	0.052	0.120	0.027	6	0.999	0.050	0.233	0.170
5000	100	50	6	0.925	0.051	0.216	0.080	6	0.755	0.048	0.548	0.216
12 500	250	50	6	1.000	0.058	0.332	0.150	6	0.996	0.047	0.726	0.559
12 500	250	50	6	0.999	0.043	0.445	0.349	6	0.992	0.059	0.777	0.527
25 000	100	250	6	1.000	0.054	0.673	0.406	6	1.000	0.057	0.934	0.750
62 500	250	250	6	0.936	0.045	0.839	0.588	5	0.992	0.050	0.942	0.924

Furthermore, the test value of the BVR-pair, rather than its p-value, is larger by quite a margin in the majority of cases (42 out of 51). For a selection of single conditions from these averages including the test values see Table E.4.

With respect to the practical use of the BVR-group and BVR-pair, the power differences in the two different types of misspecification could prove informative and can be used to identify potential model improvements. Where the BVR-group generally has a higher power to detect a missing group-level class, the BVR-pair is better able to detect missing direct effects. Since a missing class has been shown to sometimes cause only one or two BVR-group values to be significant, the conclusion could be drawn that only one or two items are problematic, rather than that an entire group-level class is missing. However, when only one item is problematic it is more likely that either the BVR-group and BVR-pair are both significant or only the BVR-pair is significant. If there is a missing class it is more likely that either both or only the BVR-group is significant. So, when only one of the two measures shows residual dependence this can be indicative of what the cause of the problem is. Of course, the wording here is deliberate in that one is more likely than the other, but not necessarily always the case.

3.5.5 Determining the Misspecified Level

Given the mutual influence of the lower- and higher-level classes, class separation and sample size, the BVR-group and BVR-pair residuals may also indicate group-level misfit, when the true problem is too few lower-level classes. Table 3.8 gives the values for the regular BVR and the BVR-group residuals when the population consists of three lower-level classes and only two are present in the estimation model. Note that the last column for the BVR values depicts the proportion of replications where one third of the BVR values are significant rather than half, thus 5 out of 15 or 15 out of 45 item pairs showing residual covariance.

It is clear that the BVR detects residual dependence between indicator items as soon as the information on the lower-level classes is sufficient, either by having a

TABLE 3.8: Power of the BVR-group and lower-level BVR to detect a missing lower-level class

Sample			Class Separation			BVR-group			BVR		
N	Groups	Group Size	Lvl. 1	Lvl. 2	G	Item 1	Min. 1	50%	Item 1	Min. 1	33%
500	50	10	L	H	2	0.040	0.246	0.002	0.130	0.908	0.100
500	50	10	L	H	3	0.328	0.970	0.126	0.248	0.998	0.002
500	50	10	H	L	2	0.056	0.440	0.002	0.610	1.000	0.732
1000	100	10	L	H	3	0.384	0.978	0.530	0.082	0.994	0.198
1000	100	10	L	L	2	0.046	0.394	0.000	0.414	1.000	0.262
1000	100	10	H	H	3	0.562	1.000	0.906	0.316	1.000	0.896
1000	100	10	H	H	2	0.080	0.472	0.002	0.844	1.000	0.996
2500	250	10	L	L	2	0.046	0.276	0.002	0.440	1.000	0.976
2500	250	10	H	L	3	0.538	1.000	1.000	0.734	1.000	0.996
2500	50	50	L	H	3	0.544	1.000	1.000	0.152	1.000	0.456
2500	50	50	H	L	3	0.526	1.000	1.000	0.666	1.000	0.930
2500	50	50	H	H	2	0.068	0.300	0.006	0.880	1.000	1.000
5000	100	50	L	H	2	0.046	0.382	0.000	0.970	1.000	1.000
5000	100	50	H	H	3	0.596	1.000	1.000	0.550	1.000	0.952

large enough sample size, or by having well defined and separated classes. Unfortunately the lower-level residual dependence is also detected by the higher-level residuals, due to the way in which they are obtained. Ideally the latter would not occur and misfit would solely be detected on the lower level.

However, as noted by Lukočiené, Varriale, & Vermunt (2010) the most fruitful strategy in fitting multilevel LC models is assuring good fit of the lower level before making adjustments to the higher level. This is also in line with studies concerning per level fit in multilevel analysis (see e.g Yuan & Bentler, 2007), where misspecification on the higher level does not systematically affect the lower-level fit when the levels are considered separately. Therefore the BVR-group and regular BVR values are contrasted for conditions with a missing higher-level class to those with a missing lower-level class in Table 3.8. In doing so it becomes clear that, although the BVR-group does report misfit when the source of that misfit originates on the lower level, the reverse does not occur. That is, the regular BVR values are very close to nominal alpha when the misfit originates on the higher level (see Table E.5 for conditions with a missing group-level class), still allowing the location of the misfit to be identifiable. Furthermore, the average proportion of significant BVR values over all replications (not reported) is similarly close to 0.05 verifying that significant values are solely due to type I errors.

3.6 Conclusion

Inspecting the properties of the two recently developed local fit statistics BVR-group and BVR-pair shows that they work as intended in detecting different types of misfit that cause residual dependence in a multilevel LC model. They allow the level of misfit to be determined, are generally capable of identifying the problematic items,

and in combination with global fit statistics and the regular bivariate residual for the lower level allow comprehensive testing and inspection of the main assumptions and substantive goals of the model.

Nonetheless, there are several issues that should be noted. First, in situations where the measures fail to detect the residual dependencies, this can have two different reasons. In cases where there is a fairly large sample on both levels, but classes are not clearly separated in terms of conditional probabilities, the residuals themselves lack power. This is not surprising, but should be kept in mind. Both the BVR-group and BVR-pair, analogous to many other fit statistics, merely test for discrepancies between model predicted and sample observed frequencies. In situations where the classes in the population are very hard to distinguish it is likely that existing dependencies can be modeled with fewer than the true number of classes and parameters. This implies that the problem is limited in that parameter bias and classification errors in these situations will be low. However, when a weakly defined class is highly relevant from a theoretical perspective, a substantive problem will remain. In turn this does mean that the residuals can be used in an exploratory setting to see whether the nested structure needs to be taken into account.

In a few, rather exceptional, situations, class separation is primarily determined by large between-group difference on only one item. The model is then able to sufficiently approach the observed frequencies while misspecified, as it can redistribute the dependence throughout the classes. This implies that not detecting misfit does not guarantee correct parameter estimation, which brings us to an important point that cannot be stressed enough. As with any residual modification index, and despite the residuals working as intended when the data is sufficient for multilevel LC analysis, they should not be used blindly. As already discussed in Chapter 2, simply trying to reduce the residuals by addressing the area of the model they report to be problematic will lead to capitalization on chance, and will hardly ever result in finding the true population model. The residuals as they are applied here, only identify the indicator items that are generally problematic. Since the different areas of the model are intertwined, they cannot point to a given solution, as any conditional dependence may be modeled in many different ways.

For practical use the general conclusion is that the residuals do provide relevant information and can help to improve model fit, but should be used in conjunction with other available measures. Also, it should be kept in mind that these are indeed residuals that detect unmodeled dependence. The briefest summary would be that if significant values are found, something is wrong in terms of capturing dependencies. By using the BVR-group and BVR-pair residuals in conjunction with global fit measures, the regular BVR, and plausible alternative models, it is possible to determine at which hierarchical level misfit occurs, identify which indicator items prove problematic, and in most cases also point at the most parsimonious way to model the uncaptured dependence. If no significant BVR-group and BVR-pair values are found, one can be sure that the nested structure of the data is captured adequately by

the model. Yet, although this is a valid conclusion, it does not always imply that the specified model agrees with the true data generating process, meaning that evaluating and comparing alternative models may still be valuable; that is, a better fitting or substantively more sensible solution can still be found when no misfit is detected.

Finally, despite this being an extensive simulation study, several factors, such as different class sizes or the addition of covariates to the model, have not been taken into account here due to the already high computational intensiveness of the current conditions. Furthermore, the relation between the detection of misfit and actual bias in parameter estimation has not been investigated, and is a valuable avenue for future research, because currently little is known about the relation between these types of misspecification and parameter estimation.

Still, for the extensive number of factors that were considered the overall conclusion is that the measures work as intended, provided that the data are sufficient for multilevel LC analysis to be viable. Although definitely requiring further research, these results also bolster our expectation that they will work for other analyses dealing with discrete nested data as well.

Chapter 4

Local Fit in Latent Markov Models

Abstract

Latent Markov, or latent transition, analysis provides a parsimonious and interpretable way to model longitudinal, categorical, panel data where the phenomenon of interest is not directly observable. However, model fit inspection and assumption testing prove difficult in this type of model. Specifically, sparseness often inhibits the use of goodness-of-fit testing, and commonly used alternative fit measures such as the AIC and BIC are not directly suitable to test specific model assumptions. To improve misfit detection and enhance the ability to search for possible model improvements, several new local fit statistics are proposed. These statistics focus on testing the two core assumptions of the model, namely capturing the nested structure of longitudinal data and the conditional independence of measurement occasions that follows from a first order Markov recursion. Applying these new local fit statistics on two different types of data examples it is shown that they are able to distinguish between different types of misfit that is due to unmodeled dependence, and allow a more directed search for model improvements.

4.1 Introduction

Research in the social sciences is often concerned with characteristics or phenomena that are not directly observable. Examples range from well-being (Samuel, Bergman, & Hupka-Brunner, 2013) or social status (Goodman, Maxwell, Malspeis, & Adler, 2015) in sociology, the degree of depression (Foli, South, Lim, & Jarnecke, 2016) or personality types (Isler, Liu, Sibley, & Fletcher, 2016) in psychology, to stock market regimes (Dias, Vermunt, & Ramos, 2015) or financial product preferences (Paas, Vermunt, & Bijmolt, 2007) in economics. In addition, the main interest generally lies in the longitudinal development of such phenomena over time, rather than their mere description. Developments that themselves are caused by not directly observable processes.

When, as in the examples, the latent phenomenon is assumed to be categorical, latent Markov (LM) models (also referred to as hidden Markov, latent transition, or regime switching models models)[1] simultaneously allow the latent concept to be measured, as well as the latent process of change to be modeled (Van de Pol & De Leeuw, 1986; Van de Pol & Langeheine, 1990; Vermunt, Langeheine, & Böckenholt, 1999). Moreover, the latent Markov approach was originally developed to take measurement error into account for multiple measurements of a single variable (Wiggins, 1955, 1973). A favorable property that has been retained throughout several extensions.

Intuitively the general multivariate LM model can be thought of as combining latent class models (Lazarsfeld & Henry, 1968; Hagenaars & McCutcheon, 2002) with a (first order) Markov chain. Based on responses to categorical items, the respondents are classified into several distinct latent classes that aim to measure a latent, underlying, categorical phenomenon. Possible transitions that the respondents may make between these classes over time are then described using a Markov chain, whereby the classes are generally referred to as states to stress the dynamic nature. The probability of transitioning from one state to another over time describes the underlying, latent, structural process, or the development of the latent phenomenon over time. Note that this presumes longitudinal panel data to model the changes of respondents over time.

The major advantages of this model are both statistical and substantive in nature, namely statistical parsimony as well as interpretability. Due to the Markov chain assumption the current state membership is only affected by the previous state, rather than by all measurement occasions. This heavily reduces the number of required parameters. Substantively, having only this temporal relation often leads to more meaningful interpretations of what happens between measurement occasions.

However, this combination of methods does come with relatively strong assumptions in both the measurement and structural part of the model. The measurement

[1] Even though all these models have specific properties and are not identical, the terms are frequently used interchangeably. For an overview see Bartolucci, Farcomeni, & Pennoni (2014).

model assumes that the different indicators are conditionally independent given the latent variable. The structural part of the model assumes that the current state is independent from all occasions other than the previous one. That is, the state at t only depends on the state at $t-1$, which is the regular Markov forward recursion. On the one hand, these assumptions lead to the two major advantages of interpretability and parsimony. On the other hand, testing these assumptions is currently problematic and makes it hard to judge the correct application of the model.

The most apparent and foremost of these problems is that, for categorical data, the model is fitted on the contingency table containing all possible answer patterns for all measurement occasions. That is, for five binary items the contingency table for one occasion consists of 32 cells. Measured at two and three occasions this increases exponentially to respectively 1024 and 32768 cells (Collins & Lanza, 2010). That this leads to problematic levels of sparseness is not surprising, and goodness-of-fit model testing becomes troublesome.

Of course, alternative fit criteria such as the BIC and AIC can be obtained, and used to compare the fit of different models, but come with problems of their own. Firstly they are relative measures that only compare estimated models to one another, and obtaining a better fit is not indicative of an overall well fitting model. That is, the better fitting model according to an information criterion may still have an overall bad fit. Secondly, these indexes only consider the global fit of the model. In more complex models with multiple assumptions, such as a LM model, it can occur that certain violations are obscured as they average out with correctly fitting parts of the model.

To circumvent some of the problems associated with model testing using global statistics, four local fit statistics are proposed in the following that will help in detecting and locating possible sources of misfit in LM models. They aim to explicitly test some of its key model assumptions, and are easily obtainable using the parameters of an estimated LM model. They use an approach analogous to the two statistics that have been proposed for the multilevel latent class model in Chapters 2 and 3 (published as Nagelkerke, Oberski, & Vermunt, 2016, 2017), which is not surprising given that the situations where members are nested in groups and measurements nested in respondents share many characteristics.

The remainder of this article is structured as follows: In the next section the LM model is briefly introduced, in the second section model fit considerations are discussed, after which an existing local fit statistic for latent class models as well as the four new fit statistics are introduced. Subsequently the proposed statistics are applied to two data examples and the results as well as the implications of those results are discussed. The final section contains additional points of discussion and the concluding remarks.

4.2 The Multivariate Latent Markov Model

Most applications of the LM model are geared towards measuring a latent, categorical phenomenon and simultaneously estimating the longitudinal development of that phenomenon. This involves classifying respondents into states or classes based on their observed responses and estimating the probabilities of moving between states given the current state membership. In other words, measuring a latent phenomenon based on observed variables, and modeling the process of change of this latent phenomenon over time.

Three parameter sets make up the core of the LM model. Firstly, similar to a latent class model, there are the item response probabilities. These describe the probability of a respondent giving a certain response to a specific item at one measurement occasion, conditional on their class membership at that time. Secondly there are the initial latent state prevalences describing the probability of belonging to a certain state at the first measurement occasion. Thirdly there are the transition probabilities of moving from one state to another between occasions (Collins & Lanza, 2010).

Combined these sets of parameters make up the model equation. The response of individual i at time t to item k is denoted as y_{itk}, with a total of N individuals, T measurement occasions, and K categorical items with R_k response categories. The vector of responses of individual i at time t is denoted $\mathbf{y}_{it}$ with $\mathbf{r}_{it}$ referring to one particular pattern, and $\mathbf{y}_i$ and $\mathbf{r}_i$ referring to the concatenation of those vectors over t. Based on the responses at each occasion a respondent can be classified into a state denoted s out of S total states on the latent variables η_t. The initial state membership at $t = 1$ is referred to as η_1. Conditional independence between the indicator items k given the latent state membership at t is assumed in the measurement model. Further assuming conditional independence of the latent states across measurement occasions t except for $t - 1$ and $t + 1$, the first-order Markov assumption, the model can be expressed as:

$$P(\mathbf{y}_i = \mathbf{r}_i) = \sum_{s_1=1}^{S} \cdots \sum_{s_T=1}^{S} P(\eta_1 = s_1) \left[\prod_{t=2}^{T} P(\eta_t = s_t | \eta_{t-1} = s_{t-1}) \right] \left[\prod_{t=1}^{T} \prod_{k=1}^{K} P(y_{itk} = r_{tk} | \eta_t = s_t) \right]. \quad (4.1)$$

The last element of Equation 4.1 is analogous to a latent class model, whereby the response probabilities are conditional on current state membership. This is essentially the measurement part of the model, where the individual responses determine state membership, and as such describe the relation between the observed indicator items and the latent variable at one occasion. The latent state membership at t results from the structural part of the model that describes the latent process, whereby the initial state membership probabilities ($\eta_1 = s_1$) are the prevalence or size of the states at $t = 1$. These are multiplied by the product of all transition probabilities, which

describe the probability of switching state membership between measurement occasions. Intuitively, the current state membership leading to the expected response frequencies is obtained by starting with the unconditional initial membership probability and subsequently considering all possible transitions between timepoints 1 and t.

The central assumptions mentioned previously are clearly separated here. In the measurement part of the model the assumption is that the items are conditionally independent given the latent variable, because it is assumed that the covariance between the items is wholly caused by the latent phenomenon; given the states, the items within each measurement occasion should be independent. In the structural part of the model the assumption is that the latent variables are conditionally independent between measurement occasions; conditional on the first order Markov chain that describes the longitudinal change in states between occasions, the latent states are assumed independent. Combining these gives an assumed conditional independence of the indicator items between and within measurement occasions. The Markov chain here implies that the latent process is without memory. That is, it holds that $P(\eta_3 = s_3|\eta_2 = s_2, \eta_1 = s_1) = P(\eta_3 = s_3|\eta_2 = s_2)$ (Vermunt, Langeheine, & Böckenholt, 1999).

Note that the more general form of the model is not often used in practice, and several additional constraints are applied to reduce the number of parameters and improve the interpretability. As formulated in Equation 4.1 the transition probabilities as well as the state definitions may change between each measurement occasion. This means that the S different states can have a substantially different meaning at each measurement occasion, impeding the interpretation of the types of subgroups that can be distinguished. Transitions from and to a state are then even more difficult to interpret, as people transition between states that have a completely different meaning over time.

In order to avoid this, one additional assumption is that of measurement invariance, whereby the state definitions are considered to be fixed by equating the item-response probabilities across occasions. As a result the transition probabilities are freely estimated, but the meaning of the states does not change over time, allowing for a far more sensible interpretation of transitions from and to states that carry more substantive weight. When measurement invariance does not hold for all indicator items, or only for certain blocks of measurement occasions, the assumption can also be applied partially.

A second additional restriction is that of homogeneous transitions, whereby the transition probabilities are assumed to be fixed over time. That is, the transitions between latent states are time invariant and identical for all pairs of adjacent occasions. This assumption mainly serves to reduce the number of parameters and simplify the substantive interpretations by not having to consider each transition separately. Of course, in cases where transition probabilities differ over time heterogeneous transitions are required to properly describe the data and the assumption needs to be

partially relaxed, or not be made at all.

In other respects, it may be useful to extend parts of the model. For example, in cases where the transition probabilities are freely estimated respondents with an identical response pattern will have identical transition probabilities. This may not be realistic as respondents can obviously differ in many other respects. Conditioning all the terms in Equation 4.1 on a covariate (Vermunt, Langeheine, & Böckenholt, 1999; Bartolucci, Pennoni, & Francis, 2007) allows respondents with different values for that covariate, for example men and women, to have different transition probabilities and follow a different longitudinal development. Moreover, the constraint of equal state definitions over time as discussed above may be too harsh even after controlling for covariates and respondents may still differ in terms of their transitions. This can be relaxed by adding a second mixture component to the model, which can be thought of as an additional level of classification that allows differing state definitions and longitudinal trajectories (Van de Pol & Langeheine, 1990).

4.3 Model Misfit & Residual Dependence

In terms of the observed data, there are a number of issues relating to how well these models fit. A well fitting LM model entails correctly reproducing the item distributions at each measurement occasion for each individual item: The model should, on average, fit each measurement occasion. Given the nested structure of the data where measurements are nested within respondents, the response sequence for each respondent should also be adequately approximated for each item: The model should capture the observed longitudinal change of a respondent.

Figure 4.1 gives an overview of potential sources of misfit and already gives some intuition on the local fit statistics that are proposed and discussed in the following. In the first panel the assumption of conditionally independent indicator items from the measurement model is shown. Given the latent variable, it is assumed that the observed items show no residual covariance. The original bivariate residual (BVR) (Vermunt & Magidson, 2013) quantifies any remaining dependence between the indicator items for each pair of variables. Although initially developed for latent class models, it directly translates to multilevel latent class and LM models.

As the item distribution of each indicator item needs to be reproduced adequately for each measurement occasion, the observed time-variable should not affect any of the indicators at any of the occasions given the model, as shown in the second panel. A similar issue occurs in multilevel models as noted in Chapter 2, where the observed group membership should not affect the items, because that would imply that the model does not fit all the groups. The same is true here and a residual effect of time is indicative of not all between-time differences being correctly modeled. This idea can be applied to the next panel as well, where the observed unit should not affect the indicators, as this would indicate that the between respondent differences are not captured by the model.

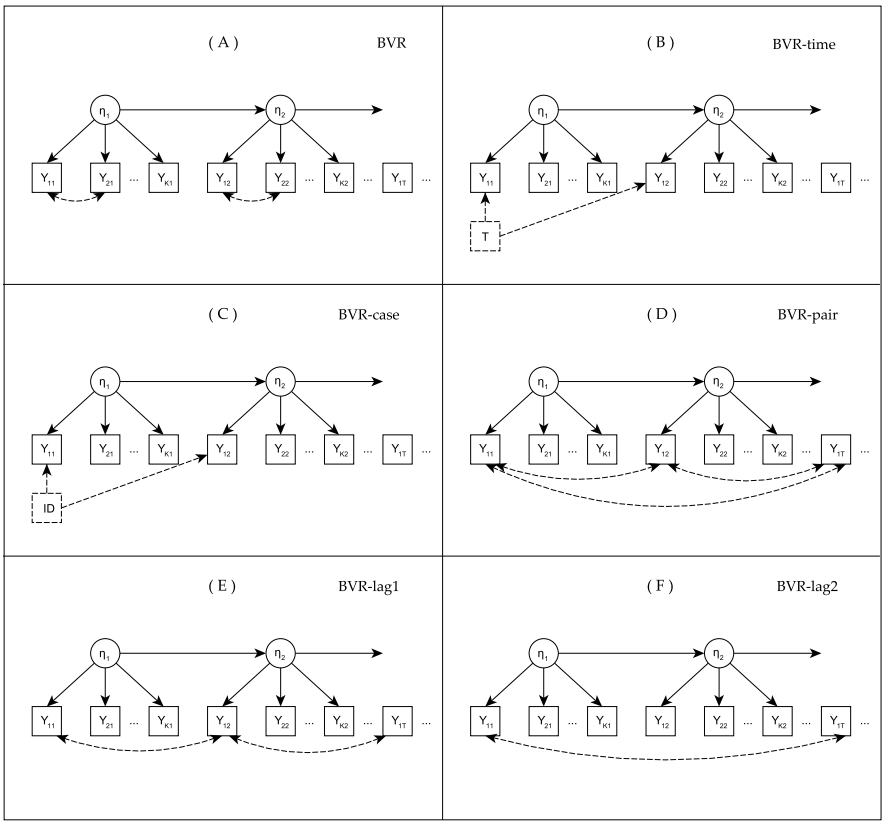

FIGURE 4.1: Overview of the BVR statistics proposed for the LM model

The final three panels of Figure 4.1 are closely related. Since the model assumes the latent phenomenon to be fully captured by the latent variable and changes in the latent phenomenon to be correctly described by the relation between the current and previous latent state, item k measured at t should be conditionally independent of any other measurement of item k. This can have several different implications. The BVR-lag1 panel shows possible residual association between adjacent measurements of an item. This would imply that despite the model aiming to capture the relation between items measured at adjacent occasions, there is residual association and the transitions over time cannot be described in enough detail. The BVR-lag2 panel shows a similar association, but here between more distant measurement occasions. When such residual dependence exists it would be an indication that the first-order Markov assumption may not hold, as more distant occasions also affect the responses to an item. The BVR-pair panel combines all possible combinations of measurement occasions, and describes residual dependence of every lag value combined. This may imply either uncaptured within-respondent dependencies, or

a violation of the first-order Markov assumption.

In summary, misfit can originate from a number of sources: (a) from residual covariance, or local dependence, between items measured at the same occasion, as well as between an item k measured at t and any other measurement of that item given the model; (b) from the model not fitting the item distributions at particular measurement occasions, whereby the manifest time variable would affect the responses to an indicator item; (c) from the structural model not describing the respondents' changes adequately, and as a combination of all these, the model not fully capturing the nested structure of the data. The local fit statistics proposed in the following are all aimed at disentangling and locating these potential sources of misfit, which may indicate ways to improve the model as well as provide valuable substantive information.

To individually test these specific aspects of a model, comparing specific combinations of expected response frequencies to the observed responses has proven a successful approach. In the regular latent class model a bivariate comparison of observed and expected frequencies is used to test for residual dependence between items that are assumed to be locally independent (Vermunt & Magidson, 2013). This bivariate residual (BVR) is also applicable to the lower level of a multilevel latent class model and similarly extends to testing the local independence of items in a LM model. Applying this logic to test whether the nested structure of the data is adequately captured by the multilevel latent class model has further proven successful in Chapters 2 and 3.

An analogous approach is proposed here, where, by constructing a Pearson-like residual, the relevant aspects of the model can be tested. One thing that must be noted is that despite all these statistics taking the form of a Pearson residual, they do not follow a chi-square distribution. This, however, is not highly problematic in practice, as p-values can be obtained relatively easily through a bootstrap procedure (Oberski, Van Kollenburg, & Vermunt, 2013; Nagelkerke, Oberski, & Vermunt, 2016, 2017).

4.3.1 Bivariate Residual (BVR)

The bivariate residual (BVR) is an already existing local fit statistic for latent class models, and only requires a slight adaptation to be used in LM models. The aim of the statistic is to detect any residual dependence between two observed indicator items given the latent variable. It does so by inspecting whether the association between a pair of variables in the observed data is properly reproduced by the estimated model. Because the indicator items are categorical variables, this is achieved by constructing a Pearson-like residual for the observed and expected response frequencies on a contingency table of two indicator items:

$$BVR_{kk'} = \frac{1}{(R_k - 1)(R_{k'} - 1)} \sum_{r=1}^{R_k} \sum_{r'=1}^{R_{k'}} \frac{(n_{rr'} - m_{rr'})^2}{m_{rr'}} \tag{4.2}$$

The observed frequencies here are simple counts of the combined responses to the two items in the data ($n_{rr'} = \sum_{i=1}^{N}\sum_{t=1}^{T} I(y_{itk} = r \;\&\; y_{itk'} = r')$). The expected frequencies ($m_{rr'}$) follow from the model as the joint probability of giving a specific response to item k and k':

$$m_{rr'} = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{s=1}^{S} P(y_{itk} = r|\eta_t = s)P(y_{itk'} = r'|\eta_t = s)P(\eta_t = s). \tag{4.3}$$

The probability of state membership $P(\eta_t = s)$ is not further conditioned or specified to keep the equation succinct here, but is the posterior probability based on the answer pattern and the transitions of the respondent.

Of course, the observed and model expected frequencies should be close to identical for the conditional independence assumption to hold. Any deviation between the two is contained in the BVR statistic as in Equation 4.2, which is then indicative of residual dependence between the pair of indicators. In practice such dependence may be resolved by explicitly modeling it and adding a covariance parameter between a pair of indicators at each measurement occasion as depicted in Figure 4.1. When it occurs between a large number of indicator items, a better solution might be to increase the number of states in the model, as the relation between the items would need a more fine-grained latent variable to be properly reproduced.

Note that the expected and observed frequencies are in principle obtained for each t and only subsequently summed. This means that, by not summing over the T measurement occasions, the statistic can be obtained for each individual measurement occasion to see whether the dependence between items primarily occurs at certain occasions or is high in general. Furthermore, the BVR statistic is not the direct Pearson-like residual, but is additionally divided $(R_k - 1)(R_{k'} - 1)$. This is done to keep the value of the statistic broadly comparable for items with different numbers of response categories, since $(R_k - 1)(R_{k'} - 1)$ is the number of non-redundant cells in the contingency table.

4.3.2 Time-variable Residual (BVR-time)

In Equation 4.1 it can be seen that the full vector of responses of an individual i is modeled. This is a vector of vectors as it were, since it contains the concatenated response patterns at each t. One step in deconstructing the global fit of the model is by inspecting on average how well the model is able to reproduce the distribution of each item k at each occasion t, rather than inspecting the overall fit across all measurement occasions and all items simultaneously.

From the model equation it follows that individual i's response to item k at occasion t can be obtained by taking the product of the state membership probabilities and the corresponding state-specific response probabilities, and summing these over

the states. The probability of a specific response is:

$$P(y_{itk} = r) = \sum_{s=1}^{S} P(\eta_t = s)P(y_{itk} = r|\eta_t = s). \tag{4.4}$$

State membership is determined by the regular Markov forward recursion, where current state membership probabilities are determined using the state membership at the previous occasion and the transition probabilities:

$$P(\eta_t = s_t) = \sum_{s_{t-1}}^{S} P(\eta_{t-1} = s_{t-1})P(\eta_t = s_t|\eta_{t-1} = s_{t-1}). \tag{4.5}$$

To inspect whether the model adequately fits the responses to a particular item at a particular measurement occasion the model expected frequencies can be compared to the observed data. Here the observed frequencies are counts of the R_k responses to item k at each t, thus $n_{ktr} = \sum_{i=1}^{N} I(y_{itk} = r)$. The model expected frequencies (m_{ktr}) can be obtained by considering the probability of a response from Equation 4.4 and summing these over the respondents at each occasion:

$$m_{ktr} = \sum_{i=1}^{N} P(y_{itk} = r). \tag{4.6}$$

Here N refers to the number of respondents observed at t, so it automatically excludes missing values. Because the interest is primarily on item level fit, in the following the actual fit statistic is obtained as a sum over T. That is:

$$BVR_{time.k} = \frac{1}{(T-1)(R_k-1)} \sum_{t=1}^{T} \sum_{r=1}^{R_k} \frac{(n_{ktr} - m_{ktr})^2}{m_{ktr}}. \tag{4.7}$$

As with the BVR, this can be thought of as constructing a cross-table for both the observed and expected frequencies and obtaining a Pearson-like residual for each item. Here, the table is not R_k by R_k, but R_k by T. In terms of such a cross-table, the number of non-redundant cells is $(R_k - 1)(T - 1)$ by which the resulting Pearson statistic is divided to make it independent of the number of categories of the variable and the number of measurement occasions.

Unfortunately the univariate marginal frequencies in the constructed observed and expected tables may differ. This would affect the value of the BVR-time, whilst not being indicative of residual time-by-variable dependence. To circumvent this issue the marginal frequencies are first equated using iterative proportional fitting (IPF). By iterating between a row and column operation this algorithm updates the cells to suit different marginal values, without affecting the cross-product ratios (Bishop, Fienberg, & Holland, 1975). Here, the expected pairwise frequencies are updated to agree with the observed marginal frequency (for an illustration see Chapter 2, Table 2.3).

A high value of the residual is then indicative of the responses to item k not being correctly reproduced at one or more measurement occasions. Not correctly reproducing an item at one or more occasions means the between-time differences are not fully captured by the model. This can be indicative of an item not fitting the latent trajectory, either because imposed measurement invariance does not hold for example, or simply because that particular item is not a good indicator for the latent phenomenon. Also note that the summation over T may be discarded, resulting in a residual for each item at each occasion. This could be substantively beneficial, for example to inspect whether items that show bad fit do so at the same measurement occasions, or to see whether the misfit for one item is consistent at all measurements or high at only a few occasions.

4.3.3 Case-variable Residual (BVR-case)

The BVR-case follows the same logic as the BVR-time, whereby a Pearson-like residual is obtained to indicate the difference between the observed and expected frequencies. Here, however, it is not the item-by-time fit that is considered, but the item-by-respondent fit. Not only should the model reproduce the item distributions at each measurement occasion, but a good fitting model also fits the response patterns of the individual respondents.

Again, the observed and model expected frequencies are required for this residual statistic to be constructed. However, a slight change to Equation 4.5 is made where not the prior, model based, state membership probabilities are used, but the posterior probabilities. The primary reason is that the individual differences are of interest here, for which person-specific prediction probabilities are needed. When using the prior, model based, probabilities the predictions would be identical for each respondent. Using the posterior state membership probabilities Equation 4.5 becomes:

$$P(y_{itk} = r) = \sum_{s_t=1}^{S} P(\eta_t = s_t|\mathbf{y}_i = \mathbf{r}_i)P(y_{itk} = r|\eta_t = s_t). \tag{4.8}$$

Note that by removing the conditioning of the state membership probability on the vector of answer patterns in $P(\eta_t = s_t|\mathbf{y}_i = \mathbf{r}_i)$ the probability of a response would indeed become identical for all respondents. To obtain the expected values we here aggregate the responses over the T occasions initially, rather than over N as done in Equation 4.6:

$$m_{ikr} = \sum_{t=1}^{T} P(y_{itk} = r) \tag{4.9}$$

The observed values are the response frequencies to one item over time, for one person. That is, for each respondent $n_{ikr} = \sum_{t=1}^{T} I(y_{itk} = r)$. Having obtained these

the BVR-case becomes:

$$BVR_{case.k} = \frac{1}{(N-1)(R_k-1)} \sum_{i=1}^{N} \sum_{r=1}^{R} \frac{(n_{ikr} - m_{ikr})^2}{m_{ikr}}. \tag{4.10}$$

The BVR-case primarily indicates whether the between-respondent differences are correctly modeled. One possible cause for misfit in this respect is that there are too few latent states and the more nuanced differences between respondents in one or more of the indicator items cannot be reflected with the current number of latent states. However, since the regular LM model tries to find one set of transitions for all respondents, it is only to a lesser extent concerned with individual differences. Misfit indicated by the BVR-case may therefore also warrant a more complex model such as a LM model with an additional mixture component. In order to better indicate whether misfit is caused by too few states or diverging temporal trajectories it can best be used in conjunction with the BVR-pair.

4.3.4 Paired-observation Residual (BVR-pair)

The BVR-time and BVR-case test the univariate reproduction of indicator items in such a way that the focus is on inspecting whether differences between measurement occasions and respondents are correctly modeled. The BVR-pair focuses on testing the within-case similarities, and thus whether the nested structure of the longitudinal data is adequately captured. That is, because respondents are observed multiple times there will be a dependence between observations at different time points that needs to be taken into account.

Testing this assumption requires a slightly different approach compared to the previous residuals, as it does not require relating two variables to one another, but requires looking at the relationship between responses within the same individual across time points. This is similar to testing the within-group dependencies in multilevel LC models, for which an intuitive solution is to create all possible pairs of respondents within each group and test whether these pairs show local independence (see also Section 2.3.3). This solution can be applied here as well, but now observations are nested within a respondent, rather than respondents within a group. The pairs that need to be created here are all possible pairs of responses to one item at the different occasions.

Creating all possible pairs of responses is done per item, per individual, and per combination of measurement occasions. Getting the observed frequency of the response-pairs becomes $n_{krr'} = \sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{x=1}^{t-1} I(y_{itk} = r \;\&\; y_{i(t-x)k} = r')$. That is, for all pairs of occasions the pair of responses to an item is identified per respondent. Intuitively, this creates an R_k by R_k table with the response pairs at t and $t-x$ per respondent. Subsequently these are summed over all respondents to give the pairwise response frequency.

The expected pairwise frequencies are obtained through the joint probability of a combination of responses, where $r'$and s' are r and s at $t-x$:

$$P(y_{itk}=r, y_{i(t-x)k}=r') = \sum_{s=1}^{S}\sum_{s'=1}^{S} P(\eta_{t-x}=s')P(\eta_t=s|\eta_{t-x}=s') P(y_{itk}=r|\eta_t=s)P(y_{i(t-x)k}=r'|\eta_{t-x}=s'). \quad (4.11)$$

The expected frequency then becomes:

$$m_{krr'} = \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{x=1}^{t-1} P(y_{i(t-x)k}=r', y_{itk}=r). \quad (4.12)$$

Obtaining $P(\eta_t = s|\eta_{t-x} = s')$, the transition probabilities between time point t and $t-x$, requires some elaboration here. Obtaining the response pair $P(y_{itk} = r, y_{i(t-x)k} = r')$ for more distant occasions namely requires the current state membership probabilities based on the response probabilities from all separate, previous measurement occasions. In case the study encompasses many occasions, what is unwanted is obtaining these probabilities in their own right, because that becomes an increasingly lengthy computation of moving over all transitions for each possible pair. A more efficient solution is to use the forward recursion and sum out the probabilities of the previous measurement occasions at each step, reducing the problem to linearly increasing with T, rather than triangularly.

To elaborate, $P(y_{i1k} = r', y_{i2k} = r)$ can be obtained from the model as:

$$P(y_{i1k}=r', y_{i2k}=r) = \sum_{s=1}^{S}\sum_{s'=1}^{S} P(\eta_1=s', y_{i1k}=r')P(\eta_2=s, y_{i2k}=r|\eta_1=s'). \quad (4.13)$$

For the next occasion $P(y_{i1k} = r', y_{i3k} = r)$ and $P(y_{i2k} = r', y_{i3k} = r)$ are needed. The current probabilities can be passed on to the next occasion. Since the joint $P(\eta_1 = s', y_{i1k} = r')$ and the transition probabilities $P(\eta_2 = s|\eta_1 = s')$ are available from the model, $P(\eta_2 = s, y_{i1k} = r')$ can be obtained as $\sum_{s'=1}^{S} P(\eta_1 = s', y_{i1k} = r')P(\eta_2 = s|\eta_1 = s')$.

Using the term $P(\eta_2 = s, y_{i1k} = r')$, the cumulative probability of pairs between the current and all previous occasions can be obtained as:

$$P(y_{i3k}=r, y_{i1k}=r') + P(y_{i3k}=r, y_{i2k}=r') = \sum_{s=1}^{S}\sum_{s'=1}^{S} [P(\eta_2=s', y_{i1k}=r') + P(\eta_2=s', y_{i2k}=r')] P(\eta_3=s, y_{i3k}=r|\eta_2=s'). \quad (4.14)$$

This sequence can be continued until all possible pairs are obtained. For the next step $P(\eta_3 = s, y_{i1k} = r') + P(\eta_3 = s, y_{i2k} = r')$ are needed, which can be obtained by $\sum_{s'=1}^{S}[P(\eta_2 = s', y_{i1k} = r') + P(\eta_2 = s', y_{i2k} = r')]P(\eta_3 = s|\eta_2 = s')$. Note that the elements between square brackets also appear in Equation 4.14. So, the probabilities from the previous occasion are passed on to the next using the transition probabilities. This implies that for the next occasion $P(\eta_3 = s', y_{i1k} = r') + P(\eta_3 = s', y_{i2k} = r')$ will get $P(\eta_3 = s', y_{i3k} = r')$ added to it in order to obtain $P(\eta_4 = s, y_{i1k} = r') + P(\eta_4 = s, y_{i2k} = r') + P(\eta_4 = s, y_{i3k} = r')$.

Thus, per pair of occasions the probabilities required are available from the previous step, except for the addition of $P(\eta_{(t-1)} = s', y_{i(t-1)k} = r')$ and multiplication with $P(\eta_t = s, y_{itk} = r|\eta_{(t-1)} = s')$, which are directly obtainable from the model. Subsequently, all pairs between the fourth and previous occasions can be obtained by plugging the above sum $P(\eta_3 = s', y_{i1k} = r') + P(\eta_3 = s', y_{i2k} = r') + P(\eta_3 = s', y_{i3k} = r')$ into equation 4.14, again multiplying it by the transition probabilities and summing over the classes for occasions three and four, etcetera:[2]

$$P(y_{i4k} = r, y_{i1k} = r') + P(y_{i4k} = r, y_{i2k} = r') + P(y_{i4k} = r, y_{i3k} = r') = \sum_{s=1}^{S}\sum_{s'=1}^{S} [P(\eta_3 = s', y_{i1k} = r') + P(\eta_3 = s', y_{i2k} = r') + P(\eta_3 = s', y_{i3k} = r')]P(\eta_4 = s, y_{i4k} = r|\eta_3 = s'). \quad (4.15)$$

Using these pairwise frequencies the same framework and broadly the same interpretation as for the other BVR statistics can be maintained, and again a Pearson-like statistic can be obtained:

$$BVR_{pair.k} = \frac{1}{R_k(R_k-1)/2}\frac{1}{N/n_t}\left[\sum_{r=1}^{R_k}\sum_{r'>r}^{R_k}\frac{((n_{krr'} + n_{kr'r}) - (m_{krr'} + m_{kr'r}))^2}{m_{krr'} + m_{kr'r}} + \sum_{r=1}^{R_k}\frac{(n_{krr} - m_{krr})^2}{m_{krr}}\right]. \quad (4.16)$$

The reason Equation 4.16 looks somewhat extensive is that here the main interest is capturing the overall, time constant, dependence between observations nested within cases. To make this explicit the pairs are considered to be interchangeable over time rather than chronologically ordered. That is, similar to a multilevel nesting, the order of observations is considered arbitrary and a yes - no pair treated as equal to a no - yes pair. This leads to treating the off-diagonal elements in the observed and expected frequency tables as symmetrical, which in Equation 4.16 is achieved by summing the mirrored off-diagonal cells together. After making these

[2] An alternative could be to approximate the posterior probabilities by using a multiplication: For lag-x the posterior probabilities would be approximated by multiplying the posterior state membership probabilities with the transition probabilities when moving from t to $t + 1$ and repeating until $t + x$ is reached. However, this would be an approximation and would stray from the actual model definition.

tables symmetrical IPF is again applied to circumvent the problem of diverging marginal frequencies similar to the BVR-time. The division by N/n_t is a division by the average number of observations per respondent (T if there are no missing observations), and is used to reduce the value of the statistic when T is large.

This also ties in with the possible model adjustments that BVR-pair may point to. The BVR-pair is indicative of any residual dependence between observations nested in respondents, which results from not fully capturing the nesting in the data. Residual within-case dependence is mostly resolved by adding time constant parameters, such as a respondent-level mixture component or a covariate. The BVR-pair is, however, most insightful in combination with the BVR-case and BVR-time, as it gives additional indications which model adjustment may prove most fruitful. For example, having a good reproduction of the longitudinal structure of the data points to model improvements that are oriented to explaining more between-respondent variance.

4.3.5 Lag-1 Residual (BVR-Lag)

When describing the BVR-pair in terms of the structural process, rather than in terms of within-person dependence, it provides an aggregate summary of all residual autocorrelation for all lag combinations. By making all the pairs of observations for all respondents the sum is that of $t = 1, \ldots, T$ combined with all $x = 1, \ldots, (t-1)$. To more precisely inspect the residual autocorrelation of items between time points the last two fit statistics proposed in this chapter are the BVR-lag1 and -lag2.

Autocorrelation is often strongest between two adjacent measurement occasions, and if the longitudinal dependence is not fully captured the expectation is that the residual dependence is strongest for a brief period of lag. Furthermore, since the BVR-pair is a lag-all residual, it can be deduced that a longer period of lag is causing residual dependence when the occasions close together show little to no residual autocorrelation.

The lag-1 autocorrelation is explicitly modeled at the latent level, as can be seen from the $P(\eta_t = s_t|\eta_{t-1} = s_{t-1})$ term in Equation 4.1. As mentioned, this translates to the item level where the observed responses should, given the model, be conditionally independent. To obtain the expected frequency of the combination of responses the same approach as for the BVR-pair is used, but here the only combination that is considered is t by $t-1$:

$$P(y_{itk} = r, y_{i(t-1)k} = r') = \sum_{s=1}^{S}\sum_{s'=1}^{S} P(\eta_{t-1} = s')P(\eta_t = s|\eta_{t-1} = s') P(y_{itk} = r|\eta_t = s)P(y_{i(t-1)k} = r'|\eta_{t-1} = s'). \quad (4.17)$$

Where the expected frequency becomes:

$$m_{krr'} = \sum_{i=1}^{N} \sum_{t=1}^{T} P(y_{itk} = r, y_{i(t-1)k} = r'). \tag{4.18}$$

The observed frequency of such a combination of responses can be counted by using an indicator function similar to the one used for the BVR-pair: $n_{krr'} = \sum_{i=1}^{N} \sum_{t=2}^{T} I(y_{itk} = r \ \& \ y_{i(t-1)k} = r')$. Having obtained the model expected and manifest observed answer patterns for adjacent measurement occasions a Pearson-like residual can again be obtained:

$$BVR_{lag-1} = \frac{1}{(T-1)(R_k-1)} \sum_{r=1}^{R_k} \sum_{r'=1}^{R_k} \frac{(n_{krr'} - m_{krr'})^2}{m_{krr'}}. \tag{4.19}$$

When the lag-1 residual is high, there is residual dependence between adjacent measurement occasions. However, since the dependence between adjacent occasions is explicitly modeled, the implication is that more fine-grained transitions should be possible either by modeling more latent states, adding covariates or adding a mixture component that would allow cases to have different transition probabilities. Essentially, lag-1 as a modeled parameter is indicative of uncaptured variance between cases in their transition from one occasion to the next. Lag-2, in contrast, is more a test of the Markov assumption of the model, since it is indicative of uncaptured autocorrelation between more distant occasions. As a result a high value for this residual implies that t is not solely dependent on $t-1$. This may be resolved by, for example, allowing heterogeneous transitions to capture the latent process more precisely, or explicitly modeling the overall lag-2 process at the latent level for the specific problematic variables.

4.4 Example Application: National Youth Study

To illustrate the use of the different fit statistics two applications are briefly presented here. The first uses data from the original National Youth Survey (Vermunt, Tran, & Magidson, 2008; Elliott, Huizinga, & Menard, 1989) containing panel data on 1725 children collected from 1976 onward, starting at the ages 11 to 17 and following them into adolescence to the ages of 27 to 33. The first five measurement occasions from 1976 thru 1980 are annual, and the latter four waves are triennial, resulting in nine separate observations. To take into account the different starting ages, and the switch in observation frequency the model is specified as if the survey is conducted annually starting at age 11, treating the ages without measurement as missing. The actual number of observations is 13665.

From these data several questions on the frequency of substance use and abuse are used, including alcohol, marijuana and harddrugs. The latter category is created by combining several questions in the original survey that distinguish between

TABLE 4.1: BVR values for the 1-state latent Markov model

	Alcohol	**Marijuana**	**Drugs**
Mrj	1800.469		
Drg	593.343	2815.124	
	Higher-level BVR		
Case	3.253	3.592	3.016
Pair	653.614	880.725	541.795
	Longitudinal BVR		
Time	122.178	51.445	22.443
Lag1	608.685	736.492	547.092
Lag2	416.784	462.414	324.170

many different types of drugs amongst which heroin, PCP, hallucinogenics such as LSD, and amphetamines such as speed. To focus on the use of the residual statistics, the data here is simplified and dichotomized into three variables, namely whether or not the respondent used alcohol, marijuana, or other drugs in the past year.

By wholly ignoring the nested structure in the data and estimating a 1-state LM model, the BVR statistics give some indication of the different types of dependence that are present in the observed data, as no dependence is modeled. As can be seen in Table 4.1 all values are high. Although in these data very apparent, estimating this baseline-model to inspect whether there truly is autocorrelation (BVR-pair, -Lag1 and -Lag2) worth modeling and enough between-item dependence (BVR) to warrant a classification already provides valuable information for data in which this might not be so abundantly clear. Note that due to the different ways in which the eventual BVR values are obtained from the raw residual their values are not comparable within one model.

To illustrate the misfit information provided by the BVRs, the approach here is exploratory. Yet, in a confirmatory setting knowing from the outset that, for example, there is limited autocorrelation may already be enough to reject or confirm several substantive ideas. Furthermore, in later stages, knowing residual dependencies and problematic variables may help in redesigning future data collection or theory.

Estimating several models constrained to have homogeneous transitions and measurement invariance with an increasing number of states would lead to selecting a 7-state model according to the BIC. However, a point not touched upon previously is that selection of the number of states in LM models is challenging. Generally, even the more penalized fit statistic have a tendency to select a high number of states (Pohle, Langrock, Van Beest, & Schmidt, 2017). This is not surprising given the specificity of the many different dependencies that are modeled. That is to say, adding more states allows, for example, the dependence between all indicator variables to be modeled better and improve the fit of the model even when that dependence is not related to the latent variable in the true model. This results in a substantively wrong, and unparsimonious model. A solution, albeit a pragmatic one, is to weigh

TABLE 4.2: Profiles for the 4-, 5-, and 6-state latent Markov model, assuming homogeneous transitions and measurement invariance

	4-State				5-state					6-state					
	1	2	3	4	1	2	3	4	5	1	2	3	4	5	6
Alc	.973	.131	.974	.995	.992	.057	.976	.858	.994	.997	.979	.110	.778	.043	.995
Mrj	.032	.010	.844	.938	.040	.007	.915	.035	.939	.043	.910	.010	.041	.003	.940
Drg	.019	.001	.067	.936	.030	.006	.073	.006	.944	.028	.073	.011	.008	.000	.946
Prev.	.420	.256	.218	.106	.330	.225	.192	.148	.106	.354	.184	.122	.121	.111	.107

the substantive interpretation and meaningfulness of the number of states against the model fit. Here too the BVR statistics may be of value.

Comparing the mean definitions of the classes across time for the 4-, 5-, and 6-state model in Table 4.2 shows that the states added in the latter models have virtually no added value in substantive terms. In the 5-state model, the definition of states one and three are close to identical and substantively no different than state one in the 4-state model. The same can be said of states three and five in the 6-state model with regard to state two of the 4-state model. Adding more states continues to improve the BIC, but leads to highly similar states substantively, which points to the model fit improvement to be mainly driven by having more transition patterns. A good starting point therefore seems the 4-state model, where the states substantively are 'Alcohol only', 'No substance use', 'Alcohol and Marijuana', and 'All substances' respectively.

When we look at the BVR values for these respective models in Table 4.3 a similar pattern arises. From the 4-state model onwards, the misfit does not get resolved very effectively by adding more states. The (not reported) 6- and 7-state model do improve in terms of overall time-dependence (BVR-time), but problems resurface in not correctly modeling the lag-1 autocorrelation. Looking at the problems with the 4-state model, there is quite a strong residual dependence between the indicator variables and time (BVR-time) and within-respondent residual dependence (BVR-pair), yet the lower order autocorrelation is modeled relatively well (BVR-lag). This indicates that the problem is mainly due to the latent process over time not being captured fully. Although adding more states is a solution in its own right, there are more parsimonious options to capture the time-dependence without adding states that differ solely in terms of transition probabilities.

That is, in the 4-state model the bivariate relations between indicator variables (BVR), the between-case variation (BVR-case), and the first-order autocorrelation (BVR-lag1) all seem to be captured relatively well when looking at the reduction from both the 1-state model in Table 4.1 and the 3-state model in Table 4.3, even though some of the residuals remain significant. Given the high values and limited reduction compared to the null model of BVR-time and BVR-pair, the main problem seems to be the variable-by-time dependence for longer periods of lag.

TABLE 4.3: BVR values for the 3-, 4- and 5-state latent Markov models assuming homogeneous transitions and measurement invariance. Bootstrap p-values based on 500 iterations between parentheses

	3-state			4-state			5-state		
	Alc	Mrj	Drg	Alc	Mrj	Drg	Alc	Mrj	Drg
Mrj	0.810 (.002)			0.973 (0.000)			0.183 (.040)		
Drg	2.587 (.008)	0.229 (.222)		0.005 (0.900)	0.095 (.136)		0.001 (.930)	0.085 (.086)	
Higher-level BVR									
Case	0.247 (.306)	0.404 (.006)	1.042 (.000)	0.282 (.064)	0.256 (.020)	0.431 (.532)	0.292 (.008)	0.234 (.032)	0.445 (.402)
Pair	33.520 (.000)	45.970 (.000)	110.484 (.000)	35.857 (.000)	30.674 (.000)	12.234 (.000)	34.650 (.000)	34.236 (.000)	17.466 (.000)
Longitudinal BVR									
Time	14.115 (.000)	37.058 (.000)	9.787 (.000)	14.243 (.000)	34.976 (.000)	9.713 (.000)	15.155 (.000)	15.639 (.000)	6.479 (.000)
Lag1	0.090 (.308)	10.548 (.000)	75.008 (.000)	0.324 (.058)	2.429 (.000)	0.749 (.000)	0.144 (.170)	1.413 (.000)	0.957 (.014)
Lag2	1.469 (.020)	4.669 (.000)	43.907 (.000)	1.858 (.080)	1.706 (.016)	1.132 (.084)	0.185 (.414)	4.095 (.000)	1.792 (.038)

TABLE 4.4: BVR values for the 4-state latent Markov models with linear, quadratic and cubic time dependence. Bootstrap p-values based on 500 iterations between parentheses

	Linear			Quadratic			Cubic		
	Alc	Mrj	Drg	Alc	Mrj	Drg	Alc	Mrj	Drg
Mrj	2.945 (.000)			2.968 (0.000)			2.885 (.000)		
Drg	0.073 (.550)	0.262 (.040)		0.121 (0.466)	0.372 (.016)		0.138 (.364)	0.364 (.012)	
Higher-level BVR									
Case	0.337 (.280)	0.247 (.234)	0.535 (.298)	0.294 (.398)	0.272 (.520)	0.545 (.304)	0.290 (.370)	0.279 (.474)	0.536 (.312)
Pair	0.071 (.614)	1.465 (.012)	1.292 (.020)	13.410 (.000)	5.027 (.000)	2.680 (.000)	12.735 (.000)	4.945 (.000)	2.851 (.000)
Longitudinal BVR									
Time	6.851 (.000)	4.775 (.000)	2.608 (.000)	0.531 (.720)	0.842 (.270)	1.524 (.014)	0.596 (.556)	0.829 (.248)	1.919 (.040)
Lag1	0.967 (.004)	3.359 (.000)	3.045 (.000)	0.046 (.500)	4.058 (.000)	2.377 (.000)	0.024 (.626)	4.720 (.000)	3.004 (.000)
Lag2	0.031 (.720)	2.725 (.000)	2.063 (.006)	3.690 (.000)	3.137 (.000)	1.343 (.018)	3.762 (.000)	3.503 (.000)	2.139 (.010)

TABLE 4.5: State membership by time in the 4-state latent Markov model with a cubic age covariate

Age	S1 - Alc Only	S2 - None	S3 - Alc & Mrj	S4 - All
11	.032	.966	.001	.001
14	.316	.495	.156	.033
17	.357	.185	.324	.135
20	.384	.107	.309	.200
23	.482	.100	.235	.182
26	.562	.119	.194	.125
29	.602	.148	.161	.089
32	.625	.177	.118	.080

Substantively it also makes sense that the longitudinal process is not fully captured by a model that assumes homogeneous transitions. From the ages 11-17 to 27-33 the transition from, for example, a non-user class to the use of alcohol, marijuana and/or other drugs will not be homogeneous over time, but will be larger at later ages (Flory, Lynam, Milich, Leukefeld, & Clayton, 2004). A non-linear effect seems plausible as well, where at later ages people are increasingly likely to start using any of the substances. A cubic effect of age would allow transitions to flatten out or reverse as people stop experimenting with drugs, which seems a further realistic pattern (Chen & Kandel, 1995). In Table 4.4 the BVR values are presented for models where age is taken into account as a covariate that affects the transitions between states. The response and state membership probabilities in these classes are for all practical purposes identical to those in Table 4.2, which is why in Table 4.5 some of the model estimated age-dependent class membership probabilities are given.

The idea of non-homogeneous transitions clearly holds, given the stark reduction in BVR-time and the completely different class sizes for different ages. Table 4.4 does show that some residual dependencies surface again when allowing such transitions. Primarily, the within-person residual dependence (BVR-pair) and residual autocorrelation (BVR-lag) starkly increase when correctly modeling the time dependencies, indicating that large parts of autocorrelation and within-person similarities over time were originally modeled as time-dependent variation. Further model improvements may be to try and increase the number of classes again, or to explicitly model some of the residual dependence when there are theoretical reasons to do so.

Without expanding the model and getting into too much detail, the above illustrates that the BVR-statistics are able to distinguish between different aspects of fit. Where global fit statistics preferred the 7-state model over the 4-state model, the BVR-pair and BVR-time show that the flexibility in transitions is central to improving the fit of this model. By allowing non-homogeneous transitions, rather than more states, the global fit of the model is improved from BIC = 28694.406 for the 4-state model, to 27740.694 for the 4-state model with a cubic time variable. The 7-state model also fits slightly worse in absolute terms (-LL = 13695.814 versus 13635.578), and is less parsimonious as it has 69 parameters, compared to 63 of the time-heterogeneous transition model.

TABLE 4.6: BVR values for the 1-state latent Markov model

		Well	**Good**
Good		1909.407	
	Higher-level BVR		
Case		2.339	2.666
Pair		118.029	181.646
	Longitudinal BVR		
Time		7.079	4.668
Lag1		60.911	76.329
Lag2		30.108	43.994

4.5 Example Application: Mood Regulation

The second application uses data of 165 German students that were prompted eight times per day, for seven days, to answer several questions on their current mood to study mood-regulation. Of the four items measured (wellness, happiness, contentment and feeling good) here the focus is on feeling good and feeling well. Originally the items contain four categories ranging from, for example, very bad to very good. However, the lowest category was used so infrequently that the categories 'very bad' and 'very unwell' are merged with 'bad' and 'unwell' respectively so that a three-category variable results. For further details on the data see Crayen, Eid, Lischetzke, Courvoisier, & Vermunt (2012). To avoid too much complication the data is considered in terms of person-days, where each day of each person is treated as a case, with each prompt per day considered as an observation. This circumvents having to consider the data structure as three-level nested data, which would overly complicate the illustration. The model now considers mood-regulation per day, with 1148 observed days.

When a 1-state model is again estimated to look at the baseline dependence in the data (Table 4.6), the bivariate association between the two variables is very large as expected. However, since these variables are deliberately chosen to be substantively highly similar and the interest lies in mood regulation what is more important is that there is a strong autocorrelation and within-person dependence (BVR-lag and BVR-pair).

As an initial step Table 4.7 shows the mean model parameters over time for the 2-, 3- and 4-state models, assuming measurement invariance and homogeneous transitions. Again, as in the National Youth Study example the states that are added after the third state are substantively highly similar to one of the existing states in the 3-state model and only differ in terms of their transition probabilities. The first and second state in the 4-state model are highly similar, with the exception that respondents in state two show a greater tendency to transfer, primarily to the second state. Furthermore, the improvement in terms of capturing more of the dependence in the observed data is limited as depicted in Table 4.8.

TABLE 4.7: Profiles for the 2-, 3-, and 4-state latent Markov model, assuming homogeneous transitions and measurement invariance

	2-State		3-state			4-state			
	1	2	1	2	3	1	2	3	4
Well									
Not	.198	.001	.047	.000	.882	.048	.032	.000	.883
Quite	.787	.300	.916	.224	.118	.944	.737	.147	.117
Very	.015	.700	.037	.776	.001	.007	.231	.853	.000
Good									
Not	.181	.010	.020	.005	.914	.018	.023	.001	.912
Quite	.678	.293	.951	.197	.085	.968	.828	.046	.088
Very	.186	.706	.029	.798	.002	.013	.150	.954	.001
Prevalence	.750	.250	.655	.210	.135	.484	.230	.151	.135
Transition									
State 1	.110	.724	.703	.101	.052	.785	.079	.067	.395
State 2	.890	.276	.264	.812	.457	.042	.178	.603	.042
State 3			.033	.088	.491	.083	.672	.302	.070
State 4						.090	.072	.028	.492

In this example, the ordinary, constrained model with substantively relevant states shows almost the same problems as the National Youth Study data. The variables still show some misfit at each individual occasion (BVR-time), and the within-day dependence is not captured all that well (BVR-pair). If the same solution to better capture the time-dependence is applied and non-homogeneous transitions are allowed, however, the within-day dependence does not get resolved as shown in Table 4.9. It does result in a better estimate of the time-dependence.

A substantively more sensible approach is to include a mixture on the day-level, to incorporate the possibility of different daily mood-regulation trajectories (Crayen et al., 2012). Such a mixture allows a classification of person-days based on the transitions that are made between each of the measurement occasions. Clearly, this model captures the nested structure of the data significantly better, although fares worse on capturing the exact item distribution at every occasion as indicated by the BVR-time. From Tabel 4.10 it does appear that two types of mood-regulation can be distinguished, where some persons have a far more resilient mood over the day and others tend to switch far more frequently.

Of course, this model could again be extended, for example by taking into account the third level of nesting to see whether respondents generally have a stable mood-regulation mechanism or whether within one respondent the mechanism shows differences between days. Regardless, the BVR statistics do point towards areas of misfit that need to be addressed and allow the global fit to be improved in a fruitful way. Moreover, they allow a more finegrained control over model inspection from a substantive perspective. As noted about Table 4.9, assuming both these models are theoretically tenable, the model with time-specific transitions is better able

TABLE 4.8: BVR values for the 2-, 3- and 4-state model, assuming homogeneous transitions and measurement invariance. Bootstrap p-values based on 500 iterations between parentheses

	2-state		3-state		4-state	
	Well	Good	Well	Good	Well	Good
Good	803.384 (.000)		0.442 (.076)		0.451 (.146)	
Higher-level BVR						
Case	1.081 (.000)	1.224 (.000)	0.441 (.098)	0.306 (.072)	0.414 (.070)	0.248 (.212)
Pair	39.822 (.000)	73.068 (.000)	9.809 (.000)	21.074 (.000)	7.153 (.000)	16.163 (.000)
Longitudinal BVR						
Time	3.322 (.000)	2.068 (.000)	3.223 (.000)	2.230 (.000)	1.940 (.000)	1.436 (.028)
Lag1	17.812 (.000)	24.957 (.000)	0.119 (.142)	0.051 (.284)	0.090 (.232)	0.046 (.292)
Lag2	8.513 (.000)	15.878 (.000)	1.593 (.000)	3.518 (.080)	1.108 (.000)	2.838 (.000)

TABLE 4.9: BVR values for the 3-state model with time-specific transitions and 2-class multilevel Markov model. Bootstrap p-values based on 500 iterations between parentheses

	Time-Specific		Mixture	
	Well	Good	Well	Good
Good	0.520 (.090)		0.309 (.108)	
Higher-level BVR				
Case	2.344 (.000)	2.670 (.000)	2.375 (.000)	2.722 (.000)
Pair	7.587 (.000)	17.400 (.000)	1.967 (.000)	4.960 (.000)
Longitudinal BVR				
Time	0.476 (.622)	2.068 (.000)	3.402 (.000)	2.199 (.000)
Lag1	0.081 (.208)	0.041 (.332)	0.441 (.000)	0.127 (.010)
Lag2	1.082 (.000)	2.821 (.000)	0.713 (.000)	1.128 (.000)

to capture the item distribution at each measurement occasion, whereas the multilevel Markov model is better able to explain the variation within each day. Of course there is the possibility to further extend the model, possibly by combining the mixture component with heterogeneous transitions, but it also serves as an example that when one of these aspects is important for the research at hand, the BVR residuals allow a choice to be made on substantive grounds. As such they may provide a warning when an overall good fitting model is used in which one important aspect is problematic.

4.6 Conclusion

In LM models global fit testing can be problematic for various reasons. Moreover, even when applicable, global fit statistics may obscure local misfit and offer little information on how to improve the model. In order to explicitly test model assumptions and get a better grasp on what aspects of the data are and are not captured by the model, four new local fit statistic were introduced that give an indication, and

TABLE 4.10: Transition probabilities for the 2-class, 3-state multilevel Markov model

Initial state probabilities				
		State 1	State 2	State 3
Class 1		.673	.154	.173
Class 2		.734	.159	.107
Transitions per state				
Conditional $(t-1)$:		State 1	State 2	State 3
Class 1	State 1	.734	.147	.119
Class 1	State 2	.394	.563	.043
Class 1	State 3	.570	.066	.365
Class 2	State 1	.970	.020	.011
Class 2	State 2	.001	.993	.006
Class 2	State 3	.084	.004	.913

allow testing of the uncaptured case-by-variabe (BVR-case), time-by-variable (BVR-time), within-respondent (BVR-pair) and autocorrelational (BVR-lag) dependencies. These statistics are also implemented in the LatentGOLD 5.1 software package.

Applying the new residual statistics to two data examples it is shown that they do indeed indicate how well a model captures the respective aspects of the data, and are able to deconstruct the global misfit. In the first application a clear age dependence that was not modeled originally was indicated by the BVR-time. Including it in the model led to a better global fit. Without the BVRs, the stark reduction in misfit and the theoretical validity of the model would probably have lead to this model being accepted as a good fitting model. However, it is shown that there may still be residual within-person dependence and autocorrelation that needs to be addressed. In the second application it is further shown that the BVRs may be of substantive value, as two different models are shown to capture two different aspects of the data better, which may both be valid for different research questions, or lead to diverging additional studies.

Of course, as was already noted for the BVR-pair and BVR-case in Chapters 2 and 3, these are residual statistics that provide an indication of aspects of the observed data that are not captured by the model. Blindly using these statistics in order to reduce them to an acceptable level does not give any guarantee of finding the population model, and leads to capitalization on chance. They do not indicate a solution, or a dependence that should be modeled in any specific way, they are merely a description of where the model expectations diverge from the observed data.

An aspect that was not touched upon is that of the comparability of the different BVR values. The way in which they are currently constructed means that their test value gives little information in its own right. Of primary importance are the reduction of the value compared to wholly ignoring the dependence, that is, the part of the dependence in the observed data that is modeled, and the p-value based on a parametric bootstrap. Further research may focus on ways to make the residual

values comparable both within models and between samples, for example by considering the average contribution to the residual per respondent and indicating a more relative or calibrated value for the residual dependence.

Finally, note that these residuals are quite flexible. Although the BVRs are here introduced for LM models, by replacing the expected values from that of any other model dealing with longitudinal categorical data (for example a mixture growth model) these statistics can be obtained in the same way. This flexibility also extends to the type of information required. Here the measures are constructed to provide the most valuable information for LM models as we see them generally applied. However, by not aggregating BVR-pair over all possible lag values for example, a residual can be obtained for every lag-distance in cases where the autocorrelation is expected to be particularly large or problematic; by not summing BVR-time over T a residual can be obtained for each occasion to, for example, inspect possible regime shifts; or by not aggregating BVR-case over N, problematic cases can be discerned. These would provide similarly intuitive information that are a vast improvement over simply having a global fit index.

Chapter 5

An Alternative Bootstrap-based Approach to Assessing Model Fit in Multilevel Latent Class Models

Abstract

In this chapter the alternative resampling approach proposed by Van Kollenburg (2017) is applied to the multilevel latent class model in order to obtain p-values for the local fit statistics BVR-group and BVR-pair. Because the approach relies on statistics that can be directly obtained from the data, these two statistics of interest cannot be used directly, because they presume a model to be estimated in the resampling process. By considering which association the BVR residuals aim to capture the residual of, two alternative chi-square statistics, BVA-group and BVA-pair, were formulated that give approximately the same information about the data. By resampling data sets from the model as would commonly be done for the parametric bootstrap, but subsequently only computing the chi-square value on these data without re-estimating the model, the distribution of the chi-square statistics under the model is obtained. When the chi-square statistic from the observed data fits nicely into this distribution, it can be concluded that the model could be the data generating process for this particular association. The results indicate nuanced differences between the BVA and BVR statistics, most notably a lower power of the BVA statistics in conditions where one indicator item is differently related to group membership than the other variables. This, however, is not necessarily the result of a truly lower power of the statistics, but rather the result of the BVA statistics answering a different question about the data.

5.1 Introduction

Decisions on the appropriateness of statistical models are often based on one or more model fit statistics. Many different of these fit statistics are available, that quantify how well, or how precise, the estimated model summarizes, describes, or explores the data. These statistics may be generally applicable or specific to the analysis model that is used, may measure goodness- or badness-of-fit, and may compare estimated models to one another or test one model against a particular value. Regardless of the model and the statistic at hand, in order to make this decision of whether the model shows a good fit to the observed data, it needs to be determined whether the value of a fit statistic is an extreme value.

Decisions on the statistical significance of fit statistics, and parameters in general, are often based on the p-value: the probability of finding the same or a more extreme value when assuming that the null-hypothesis holds in the population. Although the center of a heated statistical debate (Gelman & Loken, 2014), and the increased advocacy for other metrics (Wagenmakers, 2007), p-values remain a valuable tool for many researchers and are omnipresent in statistical research. Possibly, exactly because they provide an interpretable probability of when something is deemed an extreme value (Greenland et al., 2016).

The two most common types of p-value are the asymptotic and bootstrap p-values. The former is easily obtained when the asymptotic distribution of a statistic is known. For example, for a homoscedastic, normally distributed variable it is exactly known what the p-value is for a certain value that is some standard deviations from the mean. The bootstrap approach is more flexible and is often used in cases where the asymptotic distribution is unknown. With a parametric bootstrap the distribution of a statistic is built by computing the value many times on randomly resampled, generated data. By doing this hundreds or thousands of times it is possible to determine whether the statistic obtained for the original data has an extreme value, simply by seeing how often a more extreme value was found in the generated data (Efron & Tibshirani, 1993).

The problem with bootstrapping goodness-of-fit statistics is that it relies on resampling of the data and re-estimation of the model of interest on that data. For smaller statistical models this is not directly problematic, but for large, complex models such as a multilevel latent class (LC) model, very large samples, or cases where many models need to be compared it may become unfeasible in practice because of the computational intensity. This happened in the simulation study in Chapter 3 where a combination of these three problems occurred and several of the simulation conditions took multiple days to compute.

Recently an alternative resampling method was proposed (Van Kollenburg, 2017) that borrows the idea of the posterior predictive check from the Bayesian framework (Gelman, Meng, & Stern, 1996) and applies it to the maximum likelihood estimation framework. It can be thought of as a shortcut to the general parametric bootstrap

that uses data resampling from an estimated model, but eliminates the need for re-estimation of the full model. It does so by focusing on a particular aspect of the data that a researcher is interested in and only obtaining a statistic from the resampled data that quantifies that precise aspect without model estimation. A very similar approach has recently been proposed for structural equation models and applied to case and model fit diagnostics (Lee, Cai, & Kuhfeld, 2016). The authors also note that testing for residual dependence is a possibility. For example, when a model has a conditional independence assumption between two variables a translation must be made where it is understood that this independence assumption means that no direct effect between the two variables exists in the model, and that testing whether this holds could be done by looking at whether the model, despite the limitation, correctly reproduces the covariance between those variables.

By definition this approach is ideal for local fit testing, since it considers only those aspects of the data that are of interest. Furthermore, by eliminating the need for an asymptotic distribution, as well as the need for full model re-estimation it could be a very quick and efficient way to test the assumptions of complex models, such as the multilevel LC model.

The BVR-group and BVR-pair introduced in Chapters 2 and 3 test two parts of a multilevel LC, namely whether the specified model captures the within-group dependence and between-group variation. The goal in this chapter is to formulate statistics that can be computed directly from the data and quantify those two aspects that the BVRs aim to test, so that this new approach may be used as an alternative to the BVRs in cases where bootstrapping is unfeasible or computationally highly intensive. Of course, while doing this, the general approach is also evaluated in terms of how well it performs compared to the standard parametric bootstrap, albeit for only one type of model.

In the remainder of this chapter the resampling method will be discussed first, after which the local fit statistics BVR-pair and BVR-group will be reformulated to be directly obtainable from the data, and be illustrated with the application as presented in Chapter 2, as well as compared to some of the results from Chapter 3.

5.2 Resampling of Statistics

The resampling method as proposed by Van Kollenburg (2017) is highly similar to a regular parametric bootstrap at its core, and starts by estimating a statistical model of choice. In our case a multilevel LC model, but of course resampling is possible for any model applied to any data set.

After obtaining the model parameters it is then possible to generate data from the model. From the model, and the term parametric, here signify that the model parameters are used in generating the data, and the distributional assumptions of the model about the data are considered to be true. For example, the estimated covariance between two variables is used in generating the data, and if the model

assumes a variable to be normally distributed, a random sample of that variable in this framework would be drawn form a normal distribution with a certain mean and variance. Although not strictly necessary the structure of the data is generally maintained, so that the generated data has an equal number of observations and in the case of nested data the same number of groups with the same number of members.

The difference between the regular bootstrap and the new framework follows from how this randomly sampled, generated data is used. The parametric bootstrap at this stage would re-estimate the entire statistical model on the generated data in order to obtain the replicate goodness-of-fit statistic of interest. The obvious downside is that this takes time, and depending on the required precision, requires several hundreds to several thousands of model estimations to get a good idea of whether the statistic found in the observed data would constitute an extreme value when compared to other samples. To circumvent this, the suggestion is to use a statistic that can be directly computed on the data without the need of estimating the model, and thus can be computed very quickly on a large number of generated data sets. This statistic should of course be indicative of the relevant aspect or property of the data. Intuitively this may be thought of as bootstrapping only a very limited part of the estimated model, requiring far less computation.

To elaborate, consider the example of a conditional independence assumption, such as is made in factor analysis or LC models. Given the latent variable in these models, the indicator variables should have no residual covariance. Assessing this with a parametric bootstrap would consist of estimating the factor analytic model on all the generated data sets, obtaining the residual covariance between the variables for each one, and comparing the residual covariance from the generated data to that of the original model. The alternative proposed would only require the covariance between two indicator variables to be obtained from the generated data without estimating the model. This is possible because the model is assumed to be true. That is, the only covariance between the indicator variables that exists in the generated data is due to the latent variable as a common cause because no other effects are estimated, and if it is broadly equal to that in the observed data, the model thus reproduces this aspect of the observed data adequately.

Overall then, what this method does is it constructs the sampling distribution of a particular statistic under the model. Having this distribution in combination with the value of the statistic in the observed data allows an answer to the question of whether the model could be the data generating process for the observed data. If the observed statistic constitutes an extreme value, either very high or very low, when compared to what the statistic under the model is likely to be, it can be concluded that it is unlikely that the estimated model holds in the population. When the observed statistic is in the center of the distribution of that statistic under the model it can be concluded that, for this particular aspect, the model could constitute the data generating process.

More formally, a given statistic $S_{\mathbf{y}}$ that quantifies the data characteristic of interest is obtained on the observed data $\mathbf{y}$. Next, the model of choice is estimated as would normally be done for the analyses obtaining the parameter estimates $\hat{\theta}$. Based on the parameter estimates new data is resampled M times, thus resulting in M generated data sets $\mathbf{y}_{rep}^{(m)}$ conditional on $\hat{\theta}$. On each of these data sets the same statistic $S_{\mathbf{y}_{rep}}^{(m)}$ is obtained, after which the proportion $P(S_{\mathbf{y}_{rep}}^{(m)} > S_{\mathbf{y}})$ and / or $P(S_{\mathbf{y}_{rep}}^{(m)} < S_{\mathbf{y}})$ can be obtained, which is indicative of whether the statistic obtained on the generated data generally has a higher or lower value than that obtained on the observed data. For the exact procedure of doing this for a chi-square statistic using a LC model see Van Kollenburg (2017).

5.3 The Multilevel Latent Class Model

What follows is a very brief introduction of the multilevel LC model, which is generally used to simultaneously classify groups and their members based on categorical, nested data. It is expressed using two equations that both resemble the expression of a regular LC model, where one describes the lower level and classifies the members, and one describes the higher level and classifies the groups (Vermunt, 2003, 2008). For more details on the model please also refer to Chapters 2 and 3.

Given that there are N respondents, one of which is denoted i, that are all a member of one group j out of J groups, and responded to K items, one of which is denoted k, with response r_k out of R_k responses, giving y_{ijk} as one response by one respondent, then groups can be classified on the latent variable ζ_j with G categories and respondents can be classified on the latent variable η_{ij} with C categories, where one of the classes is referred to as g and c respectively. The model classifies respondents based on their full response pattern to the K items, which is denoted as the vector $\mathbf{y}_{ij}$, where one pattern is denoted $\mathbf{r}$. On the lower level of the model it is assumed that the K items are conditionally independent given both η_{ij} and ζ_j. It can then be expressed as:

$$P(\mathbf{y}_{ij} = \mathbf{r}|\zeta_j = g) = \sum_{c=1}^{C} P(\eta_{ij} = c|\zeta_j = g) \prod_{k=1}^{K} P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g). \quad (5.1)$$

Disregarding all elements referring to the group, so the conditioning on $\zeta_j = g$ and subscripts j, Equation 5.1 is a regular LC model, where respondents are classified based on their response pattern $P(\mathbf{y}_i = \mathbf{r})$. The probability of observing this pattern $\mathbf{r}$ is the probability of all individual responses r_k conditional on the latent class membership of the respondent, namely the product of $P(y_{ijk} = r_k|\eta_{ij} = c)$, and the probability of being in a class, namely $P(\eta_{ij} = c)$. Conditioning on the group-level latent classes ζ_j allows the probability of these response pattern to be affected by the group-level latent class membership.

On the higher level the idea is to classify the groups with regard to the responses of the group members. This classification is based on the entire response patterns of all the n_j members of a group, where these patterns are concatenated in the vector $\mathbf{y}_j$ per group j, and one of these combinations of all the $K \times n_j$ responses is referred to as $\mathbf{s}$. This part of the model assumes that the individual response patterns are conditionally independent given the group-level latent variable ζ_j and is expressed as:

$$P(\mathbf{y}_j = \mathbf{s}) = \sum_{g=1}^{G} P(\zeta_j = g) \prod_{i=1}^{n_j} P(\mathbf{y}_{ij} = \mathbf{r} | \zeta_j = g) \tag{5.2}$$

Here the probability of observing a particular group, defined as the combination of all the responses by all the members of that group $P(\mathbf{y}_j = \mathbf{s})$ is modeled highly similar to the lower level of the model, namely as the product of the probabilities of observing the response patterns of its members conditional on the class membership of the group $P(\mathbf{y}_{ij} = \mathbf{r}|\zeta_j = g)$ and the unconditional probability, or size, of the group-level class $P(\zeta_j = g)$. To avoid each group-level LC having its own lower-level class definitions the additional constraint $P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) = P(y_{ijk} = r_k|\eta_{ij} = c)$ is needed. This fixes the response probabilities in the lower-level classes to be identical for all group-level classes and the group-level classification is achieved by the different classes having different compositions of lower-level classes (Lukočiené, Varriale, & Vermunt, 2010).

5.3.1 Resampling in the Multilevel Latent Class Model

Parametric resampling in the multilevel LC model with categorical data is quite intuitive, since it is a probability or logit based model. Based on the maximum likelihood estimates of the model parameters on the original data, the sizes, or prevalences of the group-level classes $P(\zeta_j = g)$ are known. Using the structure of the original data in terms of the number of groups and their sizes the group-level class membership can be sampled first from this unconditional, multinomial probability for each of the groups. Next the class membership of the group members can be sampled conditional on the group-level class membership using $P(\eta_{ij} = c|\zeta_j = g)$. Having sampled these, the actual responses needed to obtain a generated data set can be sampled from the response probabilities conditional on both group- and individual-level class membership $P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g)$. This way a replicate of the original data is generated where the assumption is that the model is true.

5.4 Relevant Statistics

In chapters 2 and 3 two local fit statistics are introduced, both of which take the form of a Pearson residual to quantify two issues that are potential sources of misfit in a multilevel LC model. The issue with residual statistics is, however, that they presume a model as they quantify the discrepancy between the observed data and

the model expected values, which require the model to be estimated. This means they cannot be directly obtained from the (generated) data and are not suited to be used in this new resampling framework. What is needed then is to consider what information these statistics provide and how to reshape them to capture the same information about the model, but such that the statistics can be obtained without requiring model estimation.

5.4.1 Bivariate Group-Item Association (BVA-group)

The BVR-group residual considers whether the distribution of an indicator item is, on average, correctly reproduced for all the observed groups. That is, it creates a Pearson residual per group for one indicator item that compares the observed and expected frequency of the responses, and aggregates this over the groups to quantify the mismatch between the observed data and the model expected data between groups.

Although not directly apparent, one way to reformulate the correct reproduction of the per-group item distributions is to consider it as a dependence, or covariance problem. Observed group membership should namely, given the model, not affect the responses of its members in the situation where the between-group differences are adequately captured by the model. In this context the observed group membership is essentially considered a categorical covariate that has zero effect on its members given perfect fit of the model.

Extending this logic to come to an aggregate measure similar to the BVR-group, where the univariate item reproduction of the model to all the observed groups at once is considered, an alternative that can be directly obtained from the data is simply the covariance between the group-membership variable and the indicator variable of interest. A common way to quantify the strength of the association between two (nominal) categorical variables is the Pearson chi-square statistic, the value of which increases the higher the dependence between the variables. Here those are the indicator variable k and group-membership j:

$$\chi^2_{group} = \frac{1}{(J-1)(R_k-1)} \sum_{j=1}^{J} \sum_{r=1}^{R_k} \frac{(n_{jkr} - m_{jkr})^2}{m_{jkr}}, \tag{5.3}$$

where n_{jkr} is the frequency of respondents in group j with response r_k, and m_{jkr} is the expected frequency when assuming independence and obtaining the expected frequencies from the marginal frequencies. Doing this for the observed ($\mathbf{y}$) and each of the generated data sets ($\mathbf{y}^{(m)}_{rep}$) allows the comparison of χ^2 values. The division by $(J-1)(R_k-1)$ does not affect the results, but is mainly there to reduce the value of the statistic in cases where many groups are observed, as is done for the BVR-group residual.

This approach seems very similar to the way in which the BVR statistics are obtained. The difference is that rather than obtaining a Pearson residual quantifying

the difference between the observed and model expected frequencies, here the chi-square value is used to quantify the association between group-membership and the indicator variable. This association is obtained for the observed data, and compared to the chi-squared values from each of the generated data sets. Thus, rather than comparing the observed and model expected frequencies directly as the residual does, here a chi-square statistic is obtained for all data sets separately where the observed values are those from the data (observed and generated) and the expected values are those when assuming independence in the same data set. This approach will be referred to as resampling the bivariate assocation (BVA).[1]

In cases where the model could be the data generating process in the population the value obtained on the observed data should be an average, non-extreme, value in the distribution of chi-square values generated under the model. Note that the fact that this is a Pearson chi-square statistic does not matter in terms of its distribution, which may or may not be a chi-square distribution, what matters is that the value obtained on the data is similar to those obtained from the generated data.

5.4.2 Bivariate Pairwise Association (BVA-pair)

The other residual statistic, the BVR-pair, considers whether, on average, the within-group dependence among group members is captured by the model. Similar to the BVR-group it creates a Pearson residual, but does so on the pairwise responses of group members. Here too obtaining a similar, but directly computable alternative leads to considering an approach similar to the residual in the sense that this association can again be quantified through a Pearson chi-square statistic based on the frequency of pairs of responses, but without comparing the observed to the model expected values. What is compared are the actual values of a chi-square statistic computed for each generated data set to quantify the deviation from independence.

The residual dependence that the BVR-pair quantifies is the association between members of the same group that is not captured by the model, where these pairs of persons i and i' are all possible combinations of group members. It is then indicative of whether the response of persons i and i' are independent given the model. The translation to a directly obtainable equivalent again leads to quantifying this dependence between group members as a chi-square statistic and inspecting whether the value obtained on the observed data fits the distribution of values obtained on the generated data. For an illustration on how to get the pairwise frequencies please refer to Chapter 2.

One additional issue is one that also arises for the BVR-pair statistic, namely that there is no difference between pairs with a different order of responses. That is, taking a dichotomous item as an example, there is no substantive difference between

[1] In LatentGOLD this is achieved by using the procedure implemented for Monte Carlo simulation, but with the model of interest as population model, and the data as the estimated model. In a (multilevel) LC model the statistics in the observed data can directly be obtained by estimating a model with only one class, which equals estimating a model without latent variables.

respondent i responding 0 and i' responding 1 versus the reverse of responses 1 and 0. This is completely arbitrary, and solely results from the ordering of the data. To circumvent this order of responses impacting the value of the statistic the contingency tables are made symmetric for both the observed and generated data before obtaining the statistic. This can be incorporated into the general equation by treating the diagonal and off-diagonal elements differently:

$$\chi^2_{pairwise} = \frac{J}{N}\frac{1}{R_k(R_k-1)/2} \left[\sum_{r}^{R_k}\sum_{r'>r}^{R_{k'}} \frac{((n_{krr'}+n_{kr'r})-(m_{krr'}+m_{kr'r}))^2}{m_{krr'}+m_{kr'r}} + \sum_{r}\frac{(n_{krr}-m_{krr})^2}{m_{krr}}\right]. \quad (5.4)$$

The division by the constant $\frac{J}{N}\frac{1}{R_k(R_k-1)/2}$ does not affect the outcome, but again, as for the BVR-group, BVR-pair and BVA-group, is here mainly to reduce the value of the statistic.

5.5 Application: Speeding up the Job Variety Classification

To see how this resampling approach compares to the parametric bootstrap the application from Chapter 2 will be reproduced to see whether approximately the same conclusions will be drawn and the same model improvements seem most fruitful. This application uses data on task variety and required creativity at work for 848 employees working in 86 different teams. It contains five categorical items measuring the employees' perception of task variety that are dichotomized and recoded so that substantively a higher score indicates more variety and use of capacities. This example was originally also used to introduce the multilevel LC model by Vermunt (2003). For further information see Van Mierlo (2003) and Section 2.4.

The substantive idea behind classifying both the employees and teams in terms of their task variety and use of capacities is that these characteristics affect a broad range of job related issues, such as job satisfaction and turnover intent (Lambert, Hogan, & Barton, 2001). Despite being largely individual outcomes, the broader context of the team may shape such outcomes to quite a large extent (Liu, Mitchell, Lee, Holtom, & Hinkin, 2012), making the classification of teams a valuable substantive addition.

Because the new resampling approach does not affect any outcomes in terms of the overall model specification and fit, the starting point is the same as that in Chapter 2, where based on a large number of BIC values the model with two group-level LCs and three lower-level LCs is selected as the best fitting model. For convenience, Table 5.1 contains the BIC values that are presented in Chapter 2. Note that the sample size on which these values are based is the number of groups J, rather than respondents N as suggested by Lukočiené, Varriale, & Vermunt (2010) (see also Lukočiené & Vermunt, 2010).

TABLE 5.1: BIC values for 29 models assuming local independence of items and indirect effects of the group-level latent variable

	Lower Level Classes				
Group-level Classes	2	3	4	5	6
1	4,820	4,818	4,837	4,861	_[c]
2	4,786	***4,785***	4,799	4,482	4,844
3	4,794	4,795	4,794	4,814	4,837
4	4,802	4,806	4,808	4,826	4,850
5	4,811	4,818	4,822	4,839	4,865
6	4,820	4,831	4,838	4,857	4,881

[a] Values obtained using the number of groups J as the sample size in the BIC computation.
[b] Constraint: $P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) = P(y_{ijk} = r_k|\eta_{ij} = c)$.
[c] Unidentified.

In Table 5.2 the BVR values for the model with two group-level classes and three lower-level classes are shown, together with the p-values resulting from the parametric bootstrap and the new BVA resampling based on the chi-square value. In relation to this, note that the interpretations do differ between the parametric bootstrap and the chi-square resampling. Where the bootstrap of the residual indicates how often a value larger than the observed residual was encountered in the generated data and indicating the p-value of misfit, the new resampling method indicates how often a larger chi-square value quantifying the association between the variables is encountered. The latter implies that a resampling value of 0.500 is the best possible value, namely in half of the cases the reproduced association between the variables of interest is slightly stronger, and in half of the cases slightly weaker than the association in the observed data. Values towards 0.00 and 1.00 respectively indicate that the association in the generated data is always weaker (in none of the replications a stronger association is found), or always stronger (in none of the replications a weaker association is found) than the association in the observed data, which point to under- and overestimation of the association by the model. See also Figure 5.1, which depicts the obtained bootstrap distribution for three of the cells in Table 5.2.

The results from the two methods are largely identical, and both indicate problems with the nonrepetitive and creative variables on the higher level of the model. The only major difference is located on the lower level of the model, where the bootstrap indicates a problem with reproducing the covariance between the variation and diverse variables, yet the observed chi-square value seems to fit nicely into the distribution under the model.

Following the resampled values, the lower level of the model shows no major misfit, and the most pressing issue is the fit of the nonrepetitive variable on the higher level. One way to resolve this issue is improving the within-group model fit and the ability of the model to capture the group-level dependence by including a direct effect from the group-level latent variable to the indicator item. This relaxes the $P(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) = P(y_{ijk} = r_k|\eta_{ij} = c)$ constraint for one particular

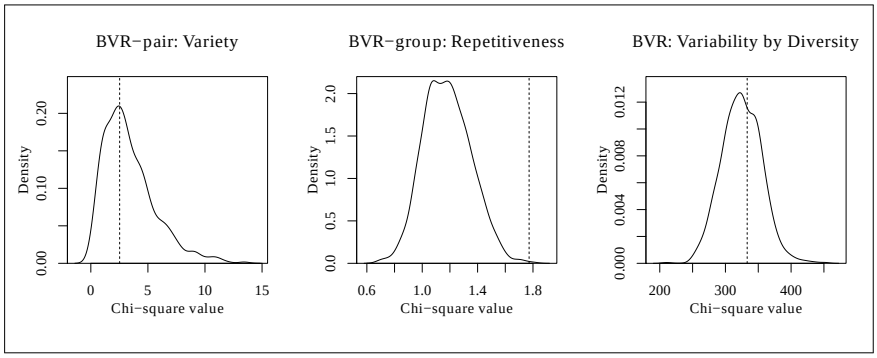

FIGURE 5.1: Distribution of Three Resampled χ^2 Statistics from Table 5.2. Dotted line indicates the value in the observed data

item and allows the response probabilities to differ between group-level classes directly, rather than only indirectly via the lower-level latent variable. When allowing this effect, which was also done in the original application in Chapter 2, the same situation occurs as previously where now all variables show problems on the higher level as depicted in Table 5.3. Both the BVR and BVA show this issue to a similar extent.

Adding a third group-level class is here the most parsimonious solution to improve the fit on the higher level. In the (not reported) three group-level class model without the direct effect on the repetitive indicator the problems with this particular item remain. In Table 5.4 the statistics for the three group-class model including the direct effect from the group-level latent variable on the nonrepetitive indicator are presented. Two issues are still being highlighted when interpreting the p-value as the significance of misfit at face value, namely that the fit of the creative tasks and use of capacities indicator variables on the higher level of the model is not ideal. Here the model is extended for the sake of illustration, but in practice this would not be advisable as the model is largely unproblematic. Firstly, because the BVA is a two-sided test the p-value for an arbitrary significance level should be twice as small. Secondly, there clearly is a multiple testing issue. Since exactly twenty p-values are presented in these tables, one significant value at an α level of .05 is to be expected on chance alone.

Choosing to continue despite the issues above, or by arbitrarily using a smaller α level, the solution is to allow a direct effect from the group-level latent variable on the indicator variables to improve fit on the higher level of the model. Allowing this effect on either one of the indicator variables does not resolve all the issues, and in Table 5.5 the values are presented for the model containing both. The profile of this model is presented in Table 5.6. This is the first model that is substantially different from the models that were found in the application using solely the bootstrap BVR statistics. Compared to following the bootstrap BVR values to improve the model, the lower-level LCs are practically identical. On the higher level, however, one large

TABLE 5.2: Residual and association values for the three class, two group-level class model. BVR, bootstrap p-values and χ^2 BVA resampling values

	Nonrepetitive			Creative			Diverse			Capacities			Variation		
	BVR	p	BVA[a]	BVR	p	BVA	BVR	p	BVA	BVR	p	BVA	BVR	p	BVA
Creative	0.763	.235	.228												
Diverse	0.248	.282	.362	0.028	.402	.532									
Capacities	0.183	.560	.696	0.359	.271	.317	0.504	.086	.768						
Variation	0.010	.743	.592	0.036	.288	.494	0.153	.021	.388	0.011	.763	.461			
BVR-group	1.586	.000	.001	1.051	.069	.034	0.788	.177	.331	1.072	.149	.043	0.816	.328	.445
BVR-pair	1.740	.000	.002	0.570	.040	.082	0.123	.320	.365	0.366	.111	.069	0.000	.966	.584

[a] The BVA is solely based on resampling and the observed χ^2 value. It is unclear what the test statistic would be, hence only the 'p-value' is presented.

TABLE 5.3: Residual and association values for the three class, three group-level class model. BVR, bootstrap p-values and χ^2 BVA resampling values

	Nonrepetitive			Creative			Diverse			Capacities			Variation		
	BVR	p	BVA	BVR	p	BVA	BVR	p	BVA	BVR	p	BVA	BVR	p	BVA
Creative	0.663	.275	.314												
Diverse	0.508	.153	.410	0.038	.346	.506									
Capacities	0.200	.567	.731	0.284	.337	.307	0.097	.290	.706						
Variation	0.097	.290	.737	0.026	.246	.534	0.233	.009	.313	0.002	.914	.493			
BVR-group	1.529	.000	.001	1.468	.002	.001	1.378	.000	.002	1.233	.042	.007	1.122	.034	.027
BVR-pair	1.587	.000	.002	1.401	.003	.082	1.397	.000	.002	0.575	.017	.011	0.263	.108	.066

and not very well defined class exists, compared to a more interpretable solution with three equally sized classes in Chapter 2, Table 2.11.

This model also has slightly worse global fit than the model as found in Chapter 2 (BIC = 4781.328 versus 4775.3), which brings about an important warning. The way in which these statistics were used in both applications is not advisable and by no means propagated as it will in almost all cases lead to capitalization on chance. In this case, blindly trying to resolve the indicated issues leads to two different substantive models, neither of which is likely to be the population model.

Despite this, the majority of the sources of misfit that are found by bootstrapping the BVR statistics are also picked up by resampling the bivariate association. One exception is the lower-level covariance that is structurally found to be problematic by the bootstrap BVR, and not by the BVA.

The major benefit of the resampling approach should be its speed, and for the final model the BVR bootstrap and BVA resampling respectively took 34.3 versus 1.2 seconds. This is already a factor 25 faster and does not take into account that the overhead computation of reading the data and preparing the output is the same in both instances and takes a non-trivial amount of time, because actually estimating this particular model on a relatively small sample is quite fast. The longer the model estimation takes, the smaller the proportion of overhead such as reading in the data set will be, and the larger the relative improvement in speed will be. Theoretically the method could be close to the number of bootstrap iterations faster, as it negates

TABLE 5.4: Residual and association values for the three class, three group-level class model: Direct effects on Repetitive and Creative indicators. BVR, bootstrap p-values and χ^2 BVA resampling values

	Nonrepetitive			Creative			Diverse			Capacities			Variation		
	BVR	*p*	BVA	BVR	*p*	BVA	BVR	*p*	BVA	BVR	*p*	BVA	BVR	*p*	BVA
Creative	0.726	.253	.320												
Diverse	0.450	.182	.444	0.014	.554	.549									
Capacities	0.205	.553	.753	0.266	.361	.372	0.367	.180	.740						
Variation	0.054	.464	.729	0.028	.276	.523	0.264	.012	.282	0.013	.792	.444			
BVR-group	0.854	.143	.183	1.057	.056	.043	0.843	.107	.222	1.050	.182	.042	0.758	.453	.480
BVR-pair	0.020	.574	.266	0.453	.067	.106	0.111	.405	.247	0.311	.118	.072	0.038	.571	.626

TABLE 5.5: Residual and association values for the three class, three group-level class model: Direct effects on Repetitive, Creative and Capacities indicators. BVR, bootstrap p-values and χ^2 BVA resampling values

	Nonrepetitive			Creative			Diverse			Capacities			Variation		
	BVR	*p*	BVA	BVR	*p*	BVA	BVR	*p*	BVA	BVR	*p*	BVA	BVR	*p*	BVA
Creative	1.067	.142	.290												
Diverse	0.215	.307	.518	0.048	.349	.504									
Capacities	0.000	.996	.582	0.413	.239	.314	0.082	.416	.634						
Variation	0.026	.553	.697	0.007	.525	.522	0.044	.073	.442	0.007	.817	.466			
BVR-group	0.952	.087	.120	0.940	.120	.279	0.862	.234	.221	0.839	.493	.502	0.845	.356	.343
BVR-pair	0.064	.345	.196	0.051	.436	.405	0.165	.230	.234	0.001	.755	.619	0.001	.910	.473

the need for model estimation. That is, 500 times quicker than a bootstrap with 500 iterations. A quick test with one of the more complex latent Markov models as used in Chapter 4, shows that the bootstrap BVR approach takes an approximate 117 minutes and the BVA resampling takes 49 seconds for 500 iterations (a factor 143 faster).

5.6 Monte Carlo Simulations

In order for the new resampling approach to be of value, it also needs to be determined whether or not the power to detect misfit can weigh up to that of the bootstrap BVR. In this section two separate simulation studies are presented. Firstly, the small simulation study from Chapter 2 is replicated in full. Secondly, a selection of the conditions from the large Monte Carlo simulation presented in Chapter 3 is re-analyzed using the BVA resampling approach.

5.6.1 Simulation: Models from the Application

To assure a realistic scenario, the final model as found in the application in Chapter 2 is used as the population model to generate data from; that is, a three class, three group-level LC model with a covariance between the variation and diverse indicator items, and two direct effects from the group-level latent variable to both the

TABLE 5.6: Profile of the three class, three group-level class model: Direct effects on Repetitive, Creative and Capacities indicators

	Group class 1 Repetitive	Group-class 2 Defined	Group-class 3 Nonrepetitive	Class 1 Diverse	Class 2 Structure	Class 3 Creative	Overall
Nonrepetitive	.228	.277	.496	.531	.119	.250	.395
Creative	.689	.304	.606	.701	.079	.807	.554
Diverse	.847	.456	.742	.961	.123	.518	.697
Capacity	.794	.486	.719	.832	.449	.351	.681
Variation	.811	.441	.703	.973	.188	.000[a]	.664
Class 1	.818	.348	.684				
Class 2	.083	.547	.200				
Class 3	.100	.105	.117				
Prevalence	.192	.227	.581	.633	.256	.111	

[a] Boundary solution

creative and the repetitive variables. The data structure is kept the same as in the original data with 853 respondents divided over 86 groups. This simulation study is performed with 1000 bootstrap or resampling iterations and 500 Monte Carlo replications. For the power computation the misspecified estimation model is a three class, two group-level LC model without any of the added parameters.

An identical approach is taken in the small simulation study presented in Chapter 2, but there are some minor differences in the power and type I error found. These are due to random fluctuation as the entire process is redone with a higher number (1000) of bootstrap iterations in order for the bootstrap BVR and resampling BVA values to be computed on the same generated data. That is, the data is generated and subsequently analyzed with both approaches. Model estimation is done twice, once for each method, in order to compare the total time required. To determine whether in both estimation steps the same model was obtained the log-likelihoods are compared. In 477 replications the same model is estimated, in the other 23 cases one of the estimations arrived at a local maximum and the likelihoods were not identical. Nonetheless, all conditions are used in the following results because the difference in likelihood on average is minute (the average difference is 0.863 on an average log-likelihood of 2470.047).

Table 5.7 shows the estimated type I error rates. That is, the proportion of Monte Carlo replications where the bootstrapped p-value exceeds significance whilst the population and estimation models are identical. Note that in this context the p-values obtained for the BVA are a two-sided test, and can either be too large or too small, thus either exceed 0.975 or are smaller than 0.025. As can be seen from the table, for both statistics the type I error rates are small, and generally adequate. A value of zero in these instances, although not expected, is largely unproblematic as long as it does not affect the power, because it only indicates that when the estimation model is identical to the population model it is never rejected.

The power of the statistics is depicted in Table 5.8, and the results are virtually identical to those found in Chapter 2 for the BVR. The BVA has lower power, something we will get back to, but does consistently detect the missing direct effect on

TABLE 5.7: Type I error for BVR-pair and BVR-group with a parametric bootstrap and for BVA resampling

	Nonrepetitive		Creative		Diverse		Capacities		Variation	
	BVR p	BVA	BVR p	BVA	BVR p	BVA	BVR p	BVA	BVR p	BVA
BVR/BVA-group	.018	.000	.026	.000	.054	.042	.028	.000	.050	0.054
BVR/BVA-pair	.034	.000	.056	.006	.052	.052	.050	.002	.064	0.042

TABLE 5.8: Type I error for BVR-pair and BVR-group with a parametric bootstrap and for BVA resampling

	Nonrepetitive		Creative		Diverse		Capacities		Variation	
	BVR p	BVA	BVR p	BVA	BVR p	BVA	BVR p	BVA	BVR p	BVA
BVR/BVA-group	.794	.750	.064	.006	.042	.040	.392	.150	.068	.028
BVR/BVA-pair	.882	.746	.078	.012	.048	.038	.410	.144	.040	.040

the nonrepetitive indicator. These power values at first glance seem quite bad for a model that is missing three parameters as well as a group-level class when compared to the population. However, the same results as in Table 5.2 are expected, where all other problems only surface after adjusting the original model. In that respect, the nonrepetitive variable that poses a problem is consistently detected as causing misfit by both methods. The second problematic indicator, creative, is not. This is not surprising when looking at the population model specified, as the logit parameters for this particular variable are an effect coded -0.186 and -0.005 (probability of .452 versus .548). Thus, although not very promising at first glance, the parameter that should be detected as causing misfit is detected with relatively high power. In terms of speed the BVA did do better, with an average duration of 7.8 for each replicate, compared to 23.3 seconds for the BVR.

5.6.2 Simulation: Synthetic Data Conditions

A selection of the conditions that are presented in Chapter 3 is re-analyzed using the BVA. It must be noted that, in contrast to the application, the BVR results in this section are those as obtained from the original Monte Carlo simulation and a new simulation is performed for the BVA statistics. This implies that the BVR and BVA statistics are obtained for the same population models and have the same misspecifications, but the generated data is not identical. That is, it can of course be expected that over the 500 Monte Carlo replications the differences in the generated data average out, but as it was impossible to retain the 25 000 data sets generated for each of the original simulation study conditions, new data is generated for the BVA computation.

Dependence in the population of these conditions is created by specifying the number of LCs, the logit parameters determining the response probabilities in both the lower- and group-level classes, and occasionally a direct effect between an indicator variable and the group-level latent variable. The general types of misspecification considered are models that do not include a parameter for one of these types

of dependence that was introduced in the population: A missing group-level LC or a missing direct effect from the group-level latent variable on the first and / or second indicator. For examples of the population profiles, see Appendix D. For an extensive discussion of the original simulation study to determine the power of the bootstrapped BVR values and its results, the reader is referred back to Chapter 3. Here the focus will be on the comparison between the BVA and BVR to detect certain types misfit.

In Table 5.9 the type I error results are presented for the 13 selected conditions, that is, conditions with the same generating and estimation models. Some aspects of this table require some elaboration. Firstly, the BVA only has one value. This is due to the synthetic nature of the data, where all groups have an identical dependence structure in the generated data and are all the same size. This causes the within-group dependence picked up by the BVA-pair, and the between-group variation picked up by the BVA-group, to be directly and unequivocally related. If the chi-square value in the generated data is higher (or lower) than the observed chi-square value for the group-item covariance, the same is true for the respondent-by-respondent dependence. This implies that the two BVA values, in this data structure, have identical power and type I error. This holds for the chi-square value per bootstrap iteration, and thus also per Monte Carlo replication.

The average proportion in Table 5.9 depicts the proportion of significant values per Monte Carlo iteration, averaged over all iterations. That is, there are as many BVR and BVA values as there are items (6 or 10). For an α of 0.05 the expectation is that, on average, five percent of all these values are significant. Similarly, the values for one particular indicator item are also expected to be significant in 0.05 of the bootstrap iterations. This is depicted for the first item. Because there is some ambiguity on the direction of the test, for the type I error the BVA is presented for both two- and one-sided tests ($.975 < x < .025$ and $x < .050$). The ambiguity is in the fact that, in principle, over-estimation of an effect is a possibility due to the independence assumptions. However, this requires the scenario that one indicator is virtually unrelated to group membership and all other indicator variables relatively strongly related to group membership in the same direction. Because the population model contains well-separated classes with indicator items all in the same direction, this is less likely than one indicator being more strongly related to group-membership than the other ones. Therefore, the expectation is that there is more gross underestimation than overestimation of parameters. As an example, see the $N = 5000$ condition in 5.9, with extremely low group-level class separation there are indeed higher type I error rates. This issue is also discussed below for power.

The results from table 5.9 are unambiguous. The type I error is controlled for both the BVR and the BVA. The low values for the BVA are unproblematic as mentioned before, since these indicate that in only a very limited number of cases the population model is rejected as the correct model. That is, they are unproblematic in their own right, unless they indicate low power to detect any misspecification.

TABLE 5.9: BVR bootstrap and BVA resampling Type I error for a selection of conditions from Chapter 3. Proportion of significant BVR and BVA values averaged over the number of indicators in the model, and the significant BVR and BVA for the first indicator item

Sample				Class Separation				Direct Effects	Avg. Proportion[a]				Item 1				Time in sec.[b]	
N	Groups	N_j	Items	Lvl.1	Lvl.2	C	G	Dir	BVR-Group	BVR-Pair	BVA 1-sided[c]	BVA 2-sided	BVR-Group	BVR-Pair	BVA 1-sided	BVA 2-sided	BVR	BVA
1000	100	10	10	0.8	0.8	3	3	2	.046	.051	.008	.002	.046	.050	.004	.000	116.1	2.9
1000	100	10	10	0.8	0.8	3	3	0	.056	.052	.005	.002	.044	.048	.002	.000	24.0	2.8
2500	50	50	10	0.7	0.7	3	3	0	.054	.050	.008	.003	.056	.056	.010	.004	84.9	3.9
2500	50	50	10	0.7	0.7	3	3	1	.049	.047	.006	.002	.068	.064	.006	.008	132.1	5.5
2500	50	50	10	0.8	0.8	3	3	0	.046	.049	.006	.000	.044	.026	.000	.000	128.8	3.1
2500	50	50	10	0.8	0.8	3	3	1	.051	.055	.001	.000	.036	.068	.000	.000	89.1	4.2
2500	250	10	6	0.8	0.8	3	3	1	.041	.047	.003	.001	.060	.060	.016	.006	125.1	3.5
5000	100	50	6	0.7	0.6	3	2	2	.047	.054	.023	.026	.040	.044	.000	.000	337.3	4.6
12 500	250	50	10	0.7	0.7	3	3	0	.047	.045	.005	.008	.028	.048	.006	.002	218.1	15.2
12 500	50	250	6	0.7	0.7	3	3	2	.065	.059	.000	.000	.052	.036	.000	.000	204.7	5.5
12 500	50	250	6	0.7	0.6	3	2	1	.057	.060	.000	.000	.048	.060	.000	.000	201.5	4.0
25 000	100	250	6	0.7	0.6	3	2	2	.039	.039	.005	.016	.044	.028	.018	.008	393.4	6.1
62 500	250	250	6	0.7	0.7	3	3	0	.053	.050	.002	.000	.040	.052	.002	.000	583.8	11.9

[a] One BVR-group, -pair, and BVA value is available per indicator item. This is the average over replications of the average proportion out of 6 or 10 values that is significant.
[b] Average time per Monte Carlo replication for 500 bootstraps in seconds. All analyses completed on identical hardware.
[c] Direction of testing is ambiguous for the BVA statistics, so the power of both one- and two-sided testing are provided.

TABLE 5.10: BVR bootstrap and BVA resampling Power for a selection of conditions from Chapter 3. Proportion of significant BVR and BVA values averaged over the number of indicators in the model, and the significant BVR and BVA for the first indicator item, and at least one significant value

Sample			Parameters						Avg. Proportion[a]			Item 1			BVA Rank[b]		At Least 1			Time in sec.[c]	
N	Groups	N_j	Items	Lvl.1	Lvl.2	C	G	Missing	BVR-Group	BVR-Pair	BVA	BVR-Group	BVR-Pair	BVA	1st	2nd	BVR-Group	BVR-Pair	BVA	BVR	BVA
1000	100	10	10	0.8	0.8	3	3	Weak direct	.049	.066	.015	.078	.266	.104	.456	.682	.402	.522	.146	85.8	2.9
1000	100	10	10	0.8	0.8	3	3	Group-class	.543	.542	.527	.560	.564	.520	-	-	1.000	1.000	1.000	28.8	2.7
2500	50	50	10	0.7	0.7	3	3	Group-class	.531	.519	.499	.554	.556	.534	-	-	1.000	1.000	1.000	152.8	3.3
2500	50	50	10	0.7	0.7	3	3	Weak direct	.076	.107	.031	.286	.490	.290	.676	.814	.578	.704	.304	146.4	5.1
2500	50	50	10	0.8	0.8	3	3	Group-class	.537	.539	.529	.564	.542	.536	-	-	1.000	1.000	1.000	49.8	3.9
2500	50	50	10	0.8	0.8	3	3	Weak direct	.061	.133	.016	.154	.826	.160	.918	.966	.460	.904	.160	142.3	3.7
2500	250	10	6	0.8	0.8	3	3	Strong direct	.081	.120	.032	.256	.214	.186	.664	.834	.410	.554	.192	125.1	3.2
5000	100	50	6	0.6	0.7	3	2	Strong direct	.073	.225	.100	.070	.988	.378	.700	.962	.990	.508	.376	337.3	5.2
12 500	250	50	10	0.7	0.7	3	3	Group-class	.525	.528	.504	.516	.518	.508	-	-	1.000	1.000	1.000	588.4	14.5
12 500	50	250	6	0.7	0.7	3	3	Strong direct	.186	.358	.082	.828	.998	.492	.978	.998	.776	.974	.492	204.7	4.1
12 500	50	250	6	0.6	0.7	3	2	Strong direct	.062	.765	.174	.072	.980	.506	.744	.808	.314	.998	.712	201.5	5.7
25 000	250	100	6	0.6	0.7	3	2	Weak direct	.171	.415	.100	.730	1.000	1.000	1.000	1.000	.804	1.000	.916	393.4	6.2
62 500	250	250	6	0.7	0.7	3	3	Group-class	.531	.525	.500	.534	.540	.502	-	-	1.000	1.000	1.000	583.8	12.0
Positive Logits for direct effects[d]																					
1000	100	10	10	0.8	0.8	3	3	Weak direct	-	-	.019	-	-	.156	.608	.806	-	-	.186	-	1.5
2500	50	50	10	0.7	0.7	3	3	Weak direct	-	-	.082	-	-	.796	.978	.994	-	-	.824	-	3.2
2500	50	50	10	0.8	0.8	3	3	Weak direct	-	-	.049	-	-	.160	.926	.986	-	-	.482	-	2.6
2500	250	10	6	0.8	0.8	3	3	Strong direct	-	-	.130	-	-	.780	.998	.998	-	-	.782	-	1.5
5000	100	50	6	0.6	0.7	3	2	Strong direct	-	-	.097	-	-	.480	.860	.976	-	-	.580	-	2.8
12 500	50	250	6	0.7	0.7	3	3	Strong direct	-	-	.037	-	-	.224	.986	1.000	-	-	.224	-	3.6
12 500	50	250	6	0.6	0.7	3	2	Strong direct	-	-	.000	-	-	.000	.710	.962	-	-	.000	-	2.8
25 000	250	100	6	0.6	0.7	3	2	Weak direct	-	-	.089	-	-	.532	1.000	1.000	-	-	.532	-	5.3

[a] One BVR-group, -pair, and BVA value is available per indicator item. This is the average over replications of the average proportion out of 6 or 10 values that is significant.
[b] Cumulative proportion of replicates in which the BVA p-value for a misspecified indicator is the largest and 2nd largest (smallest for $\alpha > .95$) even when not significant.
[c] Average time per Monte Carlo replication for 500 bootstraps in seconds. All analyses completed on identical hardware.
[d] In the BVR simulation the direct effects are negative logits, which is inconsequential the BVR p-value. To avoid artifacts when testing the BVA $x > .95$ both are presented.

In Table 5.10 the power of the statistics is depicted for different types of misspecification. Here too several remarks are required, which is mainly due to choices that are inconsequential for the BVR simulation, do matter in terms of the BVA. Firstly, all of the conditions are tested one-sided to get a fair comparison between the two approaches. It namely is the case that here the model is explicitly designed to have too few parameters, making underestimation of the covariance structures the only realistic scenario. Furthermore, direct effects from the group-level latent variable were specified to have a negative effect on the group-level separation in the original simulation study. As said, this is inconsequential for the bootstrap of the BVRs, but requires the BVA to detect the absence of the parameter through over-estimation of the dependence structure. To avoid potentially creating artificial results because of this, the models are presented with logits in both directions that are tested accordingly ($x > .95$ and $x < .05$). Note that this is different from true over-estimation as it actually is under-estimation of a negative effect that still causes dependent response patterns. Also note in this respect, that the power of significant detection of the misspecification is used as a good summary indication, but that in a normal application the right-hand side area under the density graph (see Figure 5.1) would be presented as p-value, for which, regardless of any predefined α level, extremely high and extremely low values would be noticed.

As can be concluded from the power of the BVA compared to that of the bootstrapped BVR, the BVA does not offer a replacement that detects misfit with equal reliability. The primary issue is that, as was also partly concluded for the power of the BVR statistics, the model is highly flexible in terms of redistributing dependence structures. Due to the many probabilities that are estimated, a weak dependence between group membership and one of the indicators that is not explicitly modeled is absorbed by very minor changes in the other logit parameters. Something that was also concluded from the estimated parameter change for certain types of misspecification in latent variable structural equation models (Oberski, 2014). This results in a reproduction of the covariance, quantified by the chi-square, that is considered to be non-problematic and in truth also will be relatively close to the covariance in the observed data. The BVRs are better able to detect any misspecification whereas the BVAs only in those cases where (relatively strong) dependence causes larger differences between classes for an indicator item (for an example, see the difference on the group level in Appendix F). In terms of the power to detect specific, in this case known, misspecifications this makes the BVA worse than the BVR-group and BVR-pair. In terms of answering the question whether, within the limits of chance, the model could be the data generating process, the answer is a simple 'yes', because the majority of the dependence does get modeled, and as indicated by the proportion of significant BVA values does not manifest itself elsewhere in the model. The BVA concludes that, on average, all the estimated within-group dependencies are similar enough to the observed dependencies to not be considered as misfit.

A result of the above is that the dependence of interest is modeled adequately,

and the BVA will not detect any misspecification, because the covariance it tests is not unacceptably low under the model. Further inspection of this idea is possible by inspecting the rank-order of the BVA for the indicator variable that should be reported as problematic. Namely, if a direct effect from the group-level latent variable is specified on an indicator variable, the expectation is that the covariance between group membership and that indicator is too low in the generated data. For simulation conditions that include such a direct effect, the number of times that the BVA was lowest for this indicator variable was counted. The BVA-rank columns show that the BVA does indicate the problematic indicator as the most important issue in almost all cases, yet the misspecification is simply not strong enough to detect it at an α level of .05.

In contrast to missing group-level dependencies for one single indicator variable, the power to detect it for multiple indicators in the form of a missing group-level class is on par with the BVR-group and BVR-pair statistics. Judging by the average proportion of significant BVA values, when a group-level class is missing it indicates that half of the indicators are problematic. This is due to the specification of the population: In a three group-class model the first class gets identical response probabilities for all indicators (e.g. 0.8 to score 1), the third class gets the inverse probability (0.2), and the middle class a combination where half of the indicator variables get low (0.2), and half get high (0.8) probabilities (see also Appendix A). The result is that if these three structures are forced into two classes, either the first or second half of the items in the middle class will show problems (half of the items will end up with probabilities of 0.2 in the 0.8 class or vice versa). This implies that half of the items should have a high BVA value, which is the case. Furthermore, due to its low type I error, the BVA pinpoints the problematic variables precisely given the average proportion of significant BVA values of exactly 0.5 in high power conditions.

5.7 Conclusion

In this chapter the alternative resampling approach of data statistics proposed by Van Kollenburg (2017) was applied to the multilevel LC model in order to obtain p-values for local fit statistics. Because the method relies on statistics that can be directly obtained from the observed and generated data, the BVR-group and BVR-pair statistics needed to be transformed. By considering which associations in the data these two local fit statistics aim to capture the residual of, two alternative chi-square measures were formulated that quantify approximately the same information about the data. The results in this chapter give way to a range of conclusions and topics for further research.

The BVA was first applied to the same data set used to introduce the multilevel BVR statistics with a small Monte Carlo simulation based on these data in Sections 5.5 and 5.6.1. Here the results show that the BVA and BVRs are capable of detecting largely the same misfit. One exception on the lower level is the strong residual

covariance between the variation and diverse variables, which is indicated by the BVR and not by the BVA. This has a logical explanation in that the data is resampled assuming the model to be the true model, without any re-estimation. This is unproblematic, unless the area of the model that causes the misfit is dominant in terms of model estimation, a suggestion of which can also be seen in Figure 5.1 - Frame 3. Note that this does not indicate anything about the degree of misfit, but the impact that one or more very strong dependencies may have on the estimated model parameters. Here that dependence between the two indicators is by far the largest ($\chi^2 = 333.126$ where the other bivariate associations in the data range from $\chi^2 = 16.174$ to $\chi^2 = 129.620$). Since this dependence is exactly what is modeled in a multilevel LC model, the impact that it has on the parameter estimates will be large. That is, the model will be geared towards reproducing this covariance in particular.

The bootstrap BVR does pick up on this, since it re-estimates the model and detects that the residual dependence is strong in terms of how it affects the model, and indicates that the model as a whole can be improved. As a result it suggests to explicitly include the covariance in the model to better reproduce the strong dependence, which has the added benefit of making the model far more flexible in terms of reproducing the other bivariate associations. The BVA fails to detect the issue, as it resamples under the model that is geared strongly to reproducing the dependence between that particular pair of variables, and as a result finds that the association under the model and in the observed data correspond with one another.

A similar conclusion can be drawn based on the simulation results using large, synthetic data sets in Section 5.6.2. Smaller and local dependencies are found to be easily absorbed by the large number of other parameters in the model. In these situations, due to reestimation of the model, the BVRs are able to detect the location of the residual dependence, yet the BVA does not since the observed and reproduced covariances correspond to one another. This, despite being an unforeseen result, may actually be a blessing in disguise. As noted in the results section, the BVA does not so much fail to detect misfit, it merely concludes that the reproduction of an observed bivariate association is adequate, and the subsequent conclusion is that, yes, the model could be the data generating process for this particular covariance. Where one of the largest problems with the BVR statistics is the possibility of capitalizing on chance, and merely modeling dependencies to get rid of local misfit is an extremely bad practice, the BVA statistics largely seem to circumvent this issue by answering a different question.

Now, whether the conclusion based on the BVA 'p-values' might simply be that the variable is unproblematic in terms of the group-level dependence structure requires further study. A very real possibility is namely still that the BVA statistics are simply underpowered in situations where only one, or a limited number, of indicator variables is problematic. Further studies could partly determine the difference between being underpowered and being unproblematic by combining information

on the degree of misspecification, the resulting parameter changes of that misspecification and the BVA p-value (see also Oberski, 2014; Khalid & Glas, 2016).

Lastly, in terms of detecting missing group-level classes, the BVA-group and -pair statistics perform equally well or better than their BVR counterparts. This already makes them a valuable addition to the toolbox when applying multilevel LC models. One of the most important, and simultaneously most difficult decisions in the application of these models is determining the number of (group-level) classes. Of course adding a mixture component on the group-level will often affect the global fit of the model quite drastically, but not necessarily so (Lukočiené, Varriale, & Vermunt, 2010). Moreover, obtaining a better global fit in terms of an information index such as the AIC or BIC does not mean that the additional class is a substantively valuable addition to the model.

Here the two BVA values offer important solutions because of their speed and detection rates. The intensive bootstrap procedure required for the BVR may be very inhibitive in an exploratory setting that is common for the applications of multilevel LC models. Determining the number of classes often requires many combinations with different numbers of lower- and higher-level classes to be estimated, for which performing the bootstrap may take several hours depending on the sample and data structure. Obtaining the BVA is by all practical considerations as fast as the regular estimation of the model, and despite its limitations is able to indicate whether or not the bivariate dependence structures on the lower level (Van Kollenburg, 2017) and univariate dependence structures on the higher level are decently reproduced. Furthermore, in case of a missing group-level class, the power and precision with which the BVA indicates which variables are problematic is of relevance. Here the BVA also provides substantive information on the data structure and the potential definition of the missing class due to its low type I error rate; those variables indicated as problematic are highly likely to be important indicators for the definition of additional classes. That is, an added group-level LC is likely to be substantively different from the current classes in terms of one or all of the problematic indicators.

That determining the number of components is a very difficult and impactful step in their application is generally true for finite mixture models, and the quick resampling approach may therefore also be very valuable in, for example, finite mixture item response models. In practically any other model the general idea behind this type of bootstrap alternative is also applicable, as long as the aspect of the data that affects fit, or is of substantive interest to the researcher, can be formulated in the form of a (descriptive) statistic that is directly obtainable from the data. Of course, and maybe a redundant note for the reader, these are not limited to chi-square values but can also be correlations, covariances, frequencies, means, etcetera.

Overall, despite being largely comparable in terms of the information they provide on the data, there are nuanced differences between the BVR-group and -pair compared to the BVA-group and -pair. Considering the goal with which the BVR-group and BVR-pair were originally developed, the definition of the BVA statistics

can be considered fundamentally closer to the intention of quantifying the degree with which univariate and bivariate dependencies are captured by the model. However, the differences between the two are such that all four can be used in conjunction and for different goals. In an exploratory, descriptive setting the BVAs can be used to speedily give insight in the number of mixture components, and the BVRs can be used to inspect the final model for gross assumption violations. In a predictive setting the BVAs can be a quick first step in determining the outline of the model and the BVRs can be used to enhance the final posterior probabilities.

Chapter 6

Conclusion & Discussion

The topics that this thesis covers are closely related, and as a result insights have developed on the earlier chapters after these were published as journal articles. This discussion therefore starts with briefly addressing the conclusions from the individual chapters.

In Chapter 2 the first two local fit statistics for multilevel latent class models are introduced, first both in terms of this thesis as well as to become available for the model. These BVR-group and BVR-pair residual statistics attempt to quantify the group-level dependence that is not captured by the model. By creating a residual for the within-group similarity and between-group variation, the idea was to come to a general overview of how well the model fits the group-level structure of the data, taking into account what might be considered two sides of the same coin.

In light of Chapter 5, these residuals indicate more than mere reproduction of the bivariate dependence structures for the individual indicator items, providing they are used in conjunction with a bootstrap to obtain their p-value. Through resampling and re-estimating the model, a very strong and, given the model parameters, difficult to reproduce dependence is still detected, even though the difference between the observed data and the reproduction by the model is no cause for concern. This in turn does mean that capitalization on chance will be a more pressing concern, which will be returned to later.

In Chapter 4 similar measures were introduced for the latent, or hidden, Markov model that were shown to lead to better fitting models if used as guidance in model adjustments. An extensive study into their exact properties is still required, similar to the simulation study presented in Chapter 3 for the multilevel BVR statistics. Despite the latter not considering several of the more advanced modeling options, such as multiple latent variables on the group level, or the addition of covariates, it still provides a thorough general insight into what these statistics do, and do not, detect. If similar properties hold for the Markov model the BVR statistics may also be of high added value in fields that are less concerned with the correct modeling of variance and covariance structures. In the areas of text recognition, unsupervised learning, and big data prediction, for example, the hidden Markov model has gathered a vast body of research. If indeed the BVR statistics developed for the Markov

model can indicate areas of the model with hard to reproduce dependencies, in addition to plain misfit, they might be able to improve such applications regardless of whether there is, true, narrowly defined model misfit.

Of course, for such uses several extensions and future developments will be required, and it is only in anticipation that such applications are proposed. For example, the residual needs to be extended to continuous-time measurements, as well as be made non-intrusive in model estimation. The latter was attempted in Chapter 5 by circumventing model re-estimation in a bootstrap procedure. The way in which model assumptions can be tested in this way is not only fast, but the question that is answered by considering local fit in this fashion is also very close to the definition of what a good fitting model entails. Namely, the correct reproduction of the aspects of the data that are of interest. This may be a very promising direction for future research, and currently more extensive versions of this type of resampling are considerd. For example, true to the posterior predictive check as its Bayesian origin, the general resampling scheme may offer an easy possibility to include parameter uncertainty in frequentist bootstraps. Furthermore, exactly because building the distribution of a statistic under the model and comparing it to the value found in the data is so close to the true definition of goodness-of-fit, an extension is considered to overlap the model-based and sampling distributions.

Disregarding all this for a moment and returning to the work that is presented in this thesis, there of course are several limitations and points of discussion that need to be addressed. One of these limitations pertains directly to the BVR statistics and although mentioned occasionally, it is maybe not stressed enough that these cannot determine when the model itself fails and has too little power. That is, when the sample (e.g. low N) or data structure (e.g. very low class separation) do not allow for a multilevel latent class or latent Markov model to be estimated, the value of the BVR statistics will not give any indication of this. This relates to the same shortcoming of Chapter 3, where it is impossible to make a distinction between the model not having enough power to model the true data structure and the BVR statistics having too low power to detect a failure to do so. The latter, however, is a theoretical problem as the outcome is the same.

In relation to this, it is also currently not possible to determine the degree or impact of the indicated misfit between the different BVR statistics. Most notably for the Markov model, where five different residuals are given, their test values cannot be compared directly. Yet, the test values, in their own right are the best indication of which of the indicator variables is most problematic in terms of the same residual. That is, of the same type of residual the values can be compared, where larger values indicate larger misfit. It subsequently depends on how this misfit is dealt with if the improvement in fit is proportional to the residual value. Due to the unknown, or non-existent, asymptotic distribution of the statistics the bootstrap p-values similarly cannot be used to compare the residuals. The p-values only indicate statistical significance of the residual in its truest sense, and very minor residual values can be

significant, whereas high test values may remain statistically insignificant. Whether the residuals can be standardized more, and be made more comparable within and between models is a possible future development. This may start as simple as controlling for the sample sizes in a sensible way. The reverse direction here is also possible, where it may be investigated whether the sum of all individual residuals can be used as a global or level-specific fit index, something that works remarkably well for the regular bivariate residual (Van Kollenburg, Mulder, & Vermunt, 2015).

This consideration of significance further brings up the point of multiple testing. In a model with five indicator variables, twenty BVR, BVR-group, and BVR-pair values are presented. Using the bootstrap p-values for these, it is likely that one is significant and this number increases rapidly with increasing numbers of indicators. The reason for not controlling for multiple testing is twofold. Firstly, the misfit, whether a chance effect or not, truly is a result of the observed data. Secondly, it will not prevent any issues that result from multiple testing in much the same way that arbitrarily raising or lowering the alpha level would. In the social and health sciences that provide the main background to the work presented blindly following fit indexes is what became known as cookbook statistics, and is a prevalent bad practice. Yet, researchers susceptible to blindly modeling residual dependence will simply address the most problematic area or variable in the model in an attempt to improve fit. Whether or not the p-value is controlled for multiple testing does not truly matter in that respect.

Which neatly brings us to the warning that has so far been in each of the chapters, and cannot be omitted from the discussion that indeed, especially by not controlling for multiple testing, capitalization on chance is a real danger. Although the latent class and latent Markov model are often applied in an exploratory setting, the parameters included in the model must of course be theoretically warranted.

Despite all these considerations, points of discussion and shortcomings, the general outset of this project was to make complex latent variable models easier to apply, and in that this thesis has a valuable contribution. Not only have different tests been developed that assist in determining the adherence to assumptions and the correct modeling of the data, they too have been shown to be of substantive value. This is especially true when considering that virtually all decisions on the number of mixture components in multilevel latent class and auto-correlation parameters in latent Markov models are currently based on (an adaptation of) the AIC or BIC. Fortunately, the BVR statistics presented are not only easily usable, they are also easily implementable (all are already part of the LatentGOLD software package) as they only require bivariate (categorical) expected frequencies from their respective models and a simple bootstrap procedure.

Appendix A

Chapter 2: Latent GOLD Syntax

The full syntax is given for the first model, only the equations and latent variables change thereafter. The high number of starting value sets and EM and NR iterations were not required for these models, but used for convenience as not to have to revisit and tweak the values.

First model, 3-class and 2-group-class without covariances or direct effects:

```
options
   maxthreads=all;
   algorithm
tolerance=1e-100 emtolerance=0,0001 emiterations=250000 nriterations=5000;
   startvalues
      seed=0 sets=500 tolerance=1e-005 iterations=500;
   bayes
      categorical=0 variances=0 latent=0 poisson=0;
   montecarlo
      allchi2 seed=0 sets=0 replicates=500 tolerance=1e-008;
   quadrature
nodes=10;
   missing
excludeall;
   output
parameters=effect  betaopts=wl standarderrors profile probmeans=posterior
 bivariateresiduals estimatedvalues=model iterationdetails;

variables
   groupid
team;
   dependent
w_rep nominal, w_cre nominal, w_div nominal, w_cap nominal, w_var nominal;
   latent
      GClass  group nominal 2,
      Cluster nominal 3;

equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
```

3-class, 2-group-class model, covariance between Variation and Diverse:

```
equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

3-class, 2-group-class model, covariance between Variation and Diverse and direct effect from group-level latent variable on Repetitive:

```
equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster + GClass;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

3-class, 3-group-class model, covariance between Variation and Diverse and direct effect from group-level latent variable on Repetitive:

```
Variables
   ...
   latent
      GClass  group nominal 3,
      Cluster nominal 3;

equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster + GClass;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

3-class, 3-group-class model, covariance between Variation and Diverse and direct effects from group-level latent variable on Repetitive and Creative:

```
equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster + GClass;
   w_cre <- 1 + Cluster + GClass;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

Appendix B

Chapter 2: Survey Questions

The survey questions are part of the Questionnaire on the Experience and Assessment of Work [NL: Vragenlijst beleving en beoordeling van de arbeid (VBBA)].

Repetition - In your work, do you repeatedly have to do the same things?

Creativity - Does your work require creativity?

Diversity - Is your work varied?

Capacity - Does your work sufficiently require all your skills and capacities?

Variety - Do you have enough variety in your work?

Veldhoven, van, Marc, Theodorus F. Meijman, Jacobus P. J. Broersen, and R. J. Fortuin. 1997. Handleiding VBBA: Onderzoek naar de beleving van psychosociale arbeidsbelasting en werkstress met behulp van de vragenlijst beleving en beoordeling van arbeid. [VBBA manual: An investigation of perceptions of psychosocial workload and work stress by means of the Dutch Questionnaire on the Experience and Evaluation of Work]. Amsterdam, NL: SKB.
See also http://www.marcvanveldhoven.com/ques.html.

Appendix C

Chapter 2: Simulation Syntax

```
options
   maxthreads=all;
   algorithm
      tolerance=1e-008 emtolerance=0.001 emiterations=25000 nriterations=500;
   startvalues
      seed=0 sets=25 tolerance=1e-005 iterations=50;
   bayes
      categorical=0 variances=0 latent=0 poisson=0;
   montecarlo
      seed=0 sets=0 replicates=500 tolerance=1e-008;
   quadrature
nodes=10;
   missing
excludeall;

outfile
'Generated.sav'
simulation=1;

variables
   caseweight
freq;
   groupid
groupid;
   dependent
w_rep nominal 2, w_cre nominal 2, w_div nominal 2, w_cap nominal 2, w_var nominal 2;
   latent
G group nominal 3,
X nominal 3;

equations
G <- 1;
X <- 1 | G;
w_rep <- 1 | X + G;
w_cre <- 1 | X + G;
w_div <- 1 | X;
w_cap <- 1 | X;
w_var <- 1 | X;
w_var <-> w_div;

{
-0.095419  0.057969

-0.700146 -0.273367
-0.788581 -0.537482
-1.074115  0.699874

 0.514964  0.172265  0.529056 -0.419327  0.461674
-0.050949 -0.795015  1.228202 -0.185614 -0.005218
-0.362885 -0.138018  0.903805
-0.143365  0.471675  0.255678
 4.182160  9.321291 -3.877563
 0.448358
}
```

Appendix D

Chapter 3: Population Profiles

Displaying all possible profiles would require 128 tables, hence a selection of what a typical population would look like is included. An overview of all profiles is available at osf.io/23mp2.

TABLE D.1: Profile of the conditional probabilities for 2 by 2 classes, high by low separation (group by individual)

	G1	G2	X1	X2	Overall
V1	.380	.620	.300	.700	.500
V2	.380	.620	.300	.700	.500
V3	.380	.620	.300	.700	.500
V4	.380	.620	.300	.700	.500
V5	.380	.620	.300	.700	.500
V6	.380	.620	.300	.800	.500
X1	.200	.800			
X2	.800	.200			
Prev.	.500	.500	.500	.500	

TABLE D.2: Profile of the conditional probabilities for 2 by 3 classes, low by low separation (group by individual)

	G1	G2	X1	X2	X3	Overall
V1	.460	.407	.700	.300	.300	.433
V2	.460	.407	.700	.300	.300	.433
V3	.460	.407	.700	.300	.300	.433
V4	.593	.540	.700	.700	.300	.567
V5	.593	.540	.700	.700	.300	.567
V6	.593	.540	.700	.700	.300	.567
X1	.400	.267				
X2	.333	.333				
X3	.267	.400				
Prev.	.500	.500	.333	.333	.333	

TABLE D.3: Profile of the conditional probabilities for 2 by 3 classes, low by high separation (group by individual)

	G1	G2	X1	X2	X3	Overall
V1	.440	.360	.800	.200	.200	.400
V2	.440	.360	.800	.200	.200	.400
V3	.440	.360	.800	.200	.200	.400
V4	.640	.560	.800	.800	.200	.600
V5	.640	.560	.800	.800	.200	.600
V6	.640	.560	.800	.800	.200	.600
X1	.400	.267				
X2	.333	.333				
X3	.267	.400				
Prev.	.500	.500	.333	.333	.333	

TABLE D.4: Profile of the conditional probabilities for 3 by 3 classes, low by low separation (group by individual)

	G1	G2	G3	X1	X2	X3	Overall
V1	.580	.360	.360	.700	.300	.300	.433
V2	.580	.360	.360	.700	.300	.300	.433
V3	.580	.360	.360	.700	.300	.300	.433
V4	.640	.640	.420	.700	.700	.300	.567
V5	.640	.640	.420	.700	.700	.300	.567
V6	.640	.640	.420	.700	.700	.300	.567
X1	.700	.150	.150				
X2	.150	.700	.150				
X3	.150	.150	.700				
Prev.	.333	.333	.333	.333	.333	.333	

TABLE D.5: Profile of the conditional probabilities for 3 by 3 classes, low by high separation (group by individual)

	G1	G2	G3	X1	X2	X3	Overall
V1	.620	.290	.290	.800	.200	.200	.400
V2	.620	.290	.290	.800	.200	.200	.400
V3	.620	.290	.290	.800	.200	.200	.400
V4	.710	.710	.380	.800	.800	.200	.600
V5	.710	.710	.380	.800	.800	.200	.600
V6	.710	.710	.380	.800	.800	.200	.600
X1	.700	.150	.150				
X2	.150	.700	.150				
X3	.150	.150	.700				
Prev.	.333	.333	.333	.333	.333	.333	

TABLE D.6: Profile of the conditional probabilities for 2 by 2 classes, high by high separation (group by individual), two strong direct effects

	G1	G2	X1	X2	Overall
V1	.184	.816	.852	.148	.500
V2	.184	.816	.852	.148	.500
V3	.320	.680	.800	.200	.500
V4	.320	.680	.800	.200	.500
V5	.320	.680	.800	.200	.500
V6	.320	.680	.800	.200	.500
X1	.200	.800			
X2	.800	.200			
Tot	.500	.500	.500	.500	

Appendix E

Chapter 3: Additional Results

Additional results for Chapter 3, including the power of the BVR-pair residual when only BVR-group is reported in the text, missing strong direct effects where only weak are reported in the text, and some individual conditions where only averages are presented in the text.

TABLE E.1: Power of the BVR-pair to detect ignoring the nested structure: The last three columns respectively indicate the power to reject fit for item one, at least one item and at least half of the items

	Sample		Class Separation			BVR-pair		
N	Groups	Size	Lvl. 1	Lvl. 2	C	Item 1	Min. 1	50%
500	50	10	L	L	3	0.046	0.408	0.002
500	50	10	H	L	3	0.076	0.350	0.008
500	50	10	L	H	3	0.562	0.980	0.688
500	50	10	H	L	2	0.694	0.996	0.834
1000	100	10	L	L	2	0.388	0.990	0.358
1000	100	10	L	H	2	0.910	1.000	1.000
1000	100	10	H	H	2	1.000	1.000	1.000
2500	50	50	H	H	3	1.000	1.000	1.000
2500	50	50	L	H	2	1.000	1.000	1.000
2500	250	10	L	H	3	0.996	1.000	1.000
2500	250	10	H	H	3	1.000	1.000	1.000
5000	100	50	L	L	3	0.196	0.838	0.028
5000	100	50	H	L	3	0.588	0.974	0.754
5000	100	50	H	L	2	1.000	1.000	1.000

TABLE E.2: Power of the BVR-pair to detect a missing group-level class: The last three columns respectively indicate the power to reject fit for item one, at least one item and at least half of the items

	Sample		Class Separation				BVR-pair		
N	Groups	Size	Items	Lvl. 1	Lvl. 2	C	Item 1	Min. 1	50%
500	50	10	6	L	H	2	0.064	0.270	0.000
500	50	10	6	H	H	3	0.542	1.000	0.930
500	10	50	6	L	H	2	0.064	0.364	0.000
500	10	50	6	H	H	3	0.452	0.976	0.884
1000	100	10	6	H	L	2	0.050	0.276	0.006
1000	100	10	10	L	H	2	0.042	0.404	0.000
1000	100	10	6	L	L	3	0.152	0.678	0.068
1000	100	10	10	H	H	3	0.564	1.000	1.000
2500	250	10	6	L	L	2	0.048	0.240	0.004
2500	250	10	10	H	L	2	0.060	0.380	0.000
2500	250	10	6	H	H	2	0.078	0.440	0.004
2500	250	10	6	L	L	3	0.396	0.944	0.396
2500	250	10	10	L	H	2	0.064	0.460	0.000
2500	250	10	6	H	L	2	0.052	0.280	0.008
2500	250	10	10	H	H	3	0.556	1.000	1.000
2500	50	50	6	H	L	2	0.110	0.514	0.024
2500	50	50	6	L	H	3	0.534	1.000	1.000
2500	50	50	10	H	H	3	0.536	1.000	1.000
2500	50	50	10	L	L	3	0.556	1.000	0.968
5000	100	50	6	L	L	2	0.072	0.386	0.002
5000	100	50	10	H	L	2	0.188	0.868	0.044
5000	100	50	6	H	L	3	0.494	1.000	1.000
5000	100	50	6	H	H	2	0.784	0.980	1.000

TABLE E.3: Power to detect the absence of a strong direct effect from the group-level latent variable on the first indicator variable

Sample			Separation					Group-level Entropy		Lower-level Entropy		Power	
N	Groups	N_j	Items	Lvl.1	Lvl.2	C	G	Pop.	Model	Pop.	Model	BVR-group	BVR-pair
500	50	10	10	L	L	2	3	0.532	0.295	0.693	0.694	0.990	0.998
500	50	10	10	L	H	3	2	0.927	0.809	0.558	0.532	0.166	0.308
500	50	10	6	L	H	2	2	0.965	0.847	0.616	0.637	0.068	0.748
500	50	10	6	L	H	3	3	0.764	0.706	0.517	0.490	0.050	0.052
500	50	10	6	H	H	3	2	0.937	0.886	0.671	0.657	0.210	0.086
500	50	10	10	H	L	3	3	0.899	0.863	0.827	0.822	0.342	0.134
500	50	10	10	H	H	2	2	0.974	0.930	0.947	0.949	0.660	1.000
1000	100	10	10	L	H	2	3	0.659	0.509	0.719	0.722	0.996	1.000
1000	100	10	6	L	L	3	2	0.689	0.075	0.323	0.324	0.126	0.524
1000	100	10	6	L	L	2	3	0.530	0.267	0.530	0.530	0.986	1.000
1000	100	10	10	L	L	2	2	0.901	0.619	0.706	0.709	1.000	1.000
1000	100	10	6	H	H	2	3	0.662	0.553	0.840	0.844	0.996	1.000
1000	100	10	10	H	L	3	2	0.659	0.144	0.782	0.781	0.992	0.970
2500	250	10	6	L	H	3	2	0.837	0.644	0.425	0.383	0.084	0.204
2500	250	10	10	L	L	3	2	0.715	0.099	0.472	0.470	1.000	1.000
2500	250	10	6	L	H	2	3	0.641	0.462	0.572	0.577	0.990	1.000
2500	250	10	10	H	L	2	3	0.509	0.336	0.936	0.937	1.000	1.000
2500	250	10	6	H	L	2	3	0.933	0.916	0.750	0.737	0.256	0.214
2500	250	10	6	H	H	2	2	0.971	0.920	0.853	0.862	0.822	1.000
2500	50	50	6	L	H	2	2	1.000	1.000	0.620	0.656	0.046	1.000
2500	50	50	10	L	H	3	3	1.000	0.999	0.680	0.665	0.856	0.934
2500	50	50	6	L	L	3	3	0.977	0.928	0.484	0.457	0.636	0.746
2500	50	50	6	H	L	3	3	0.986	0.381	0.618	0.614	0.970	1.000
2500	50	50	10	H	L	2	3	0.929	0.780	0.937	0.938	1.000	1.000
5000	100	50	10	L	L	3	2	0.997	0.326	0.478	0.472	0.874	1.000
5000	100	50	6	L	H	2	3	0.982	0.910	0.580	0.603	0.838	1.000
5000	100	50	10	L	L	2	3	0.946	0.741	0.696	0.703	1.000	1.000
5000	100	50	6	H	L	2	2	1.000	0.994	0.829	0.837	1.000	1.000
5000	100	50	10	H	L	3	3	1.000	1.000	0.835	0.830	1.000	0.742
12 500	250	50	6	L	L	3	2	0.994	0.252	0.330	0.325	0.052	1.000
12 500	250	50	10	L	L	2	2	1.000	0.991	0.704	0.717	1.000	1.000
12 500	250	50	10	L	L	3	3	0.997	0.987	0.599	0.585	1.000	1.000
12 500	250	50	10	H	H	2	3	0.987	0.962	0.943	0.945	1.000	1.000
12 500	250	50	10	H	H	3	2	1.000	1.000	0.815	0.808	1.000	0.918
12 500	250	50	6	H	L	2	3	0.929	0.773	0.825	0.830	1.000	1.000

TABLE E.4: Selection of single conditions out of the averages in Table 3.7. The values of the residuals are the averages of the bootstrap mean values over replications

Sample			Class Separation					Item 1		Item 2		Item 3	
N	Groups	Size	Items	Lvl. 1	Lvl. 2	C	G	Power	Val.	Power	Val.	Power	Val.
500	50	10	10	L	L	2	2	0.998	5.010	0.040	0.044	0.078	0.167
500	50	10	10	H	L	3	3	0.118	0.198	0.050	0.088	0.064	0.143
1000	100	10	10	L	L	3	3	0.182	0.233	0.056	0.090	0.060	0.108
1000	100	10	10	H	L	3	2	0.944	2.227	0.052	0.043	0.048	0.103
2500	50	50	6	L	H	2	3	1.000	5.302	0.040	0.011	0.358	0.171
2500	50	50	6	H	L	2	2	1.000	51.77	0.048	0.011	0.080	0.102
2500	250	10	10	L	L	3	3	0.812	1.140	0.044	0.074	0.178	0.211
2500	250	10	6	H	H	2	3	1.000	18.18	0.030	0.010	0.212	0.328
5000	100	50	6	L	L	3	2	0.988	0.575	0.042	0.016	0.142	0.036
5000	100	50	10	H	L	3	3	0.700	0.448	0.054	0.019	0.246	0.214
12 500	50	250	6	L	L	3	3	0.976	0.407	0.048	0.016	1.000	5.814
12 500	50	250	10	H	L	2	3	1.000	121.8	0.042	0.009	0.150	0.123
12 500	250	50	6	L	L	2	2	1.000	21.26	0.046	0.002	1.000	1.135
12 500	250	50	6	H	H	3	3	0.970	0.599	0.042	0.004	1.000	1.220
25 000	100	250	10	L	H	2	2	1.000	272.5	0.040	0.002	1.000	1.545
25 000	100	250	6	H	H	2	3	1.000	196.6	0.058	0.003	0.998	1.982
62 500	250	250	6	L	H	3	3	1.000	3.042	0.052	0.004	1.000	0.935
62 500	250	250	6	H	L	3	3	0.960	0.913	0.044	0.004	1.000	12.22

TABLE E.5: Power of the lower-level BVR values when a group-level latent class is missing

Sample			Class Separation				BVR-group		
N	Groups	Group Size	Items	Lvl.1	Lvl.2	C	Item 1-2	Item 1-3	Item 3-4
500	50	10	6	L	H	2	0.032	0.044	0.060
500	50	10	6	H	H	3	0.052	0.040	0.054
1000	100	10	6	H	L	2	0.080	0.056	0.040
1000	100	10	10	L	H	2	0.040	0.054	0.050
1000	100	10	6	L	L	3	0.030	0.038	0.046
1000	100	10	10	H	H	3	0.044	0.052	0.058
2500	250	10	6	L	L	2	0.056	0.052	0.050
2500	250	10	10	H	L	2	0.050	0.046	0.038
2500	250	10	6	H	H	2	0.042	0.044	0.042
2500	50	50	6	H	L	2	0.050	0.036	0.050
2500	50	50	6	L	H	3	0.054	0.054	0.044
2500	50	50	10	H	H	3	0.044	0.048	0.050
2500	50	50	10	L	L	3	0.064	0.060	0.042
5000	100	50	6	L	L	2	0.060	0.052	0.032
5000	100	50	10	H	L	2	0.044	0.046	0.050
5000	100	50	6	H	L	3	0.040	0.050	0.060
5000	100	50	6	H	H	2	0.040	0.052	0.048

Appendix F

Chapter 5: Population Profiles

TABLE F.1: Profile of the conditional probabilities for 3 by 3 classes, high by high separation (group by individual). Positive direct effect from group-level latent on first variable

	G1	G2	G3	X1	X2	X3	Overall
V1	.816	.260	.133	.873	.209	.127	.403
V2	.680	.260	.260	.800	.200	.200	.400
V3	.680	.260	.260	.800	.200	.200	.400
V4	.740	.740	.320	.800	.800	.200	.600
V5	.740	.740	.320	.800	.800	.200	.600
V6	.740	.740	.320	.800	.800	.200	.600
X1	.800	.100	.100				
X2	.100	.800	.100				
X3	.100	.100	.800				
Prev.	.333	.333	.333	.333	.333	.333	

First item has a logit of 0.5108256 from the group-level latent variable

TABLE F.2: Profile of the conditional probabilities for 3 by 3 classes, high by high separation (group by individual). Negative direct effect from group-level latent on first variable

	G1	G2	G3	X1	X2	X3	Overall
V1	.489	.260	.461	.644	.209	.356	.403
V2	.680	.260	.260	.800	.200	.200	.400
V3	.680	.260	.260	.800	.200	.200	.400
V4	.740	.740	.320	.800	.800	.200	.600
V5	.740	.740	.320	.800	.800	.200	.600
V6	.740	.740	.320	.800	.800	.200	.600
X1	.800	.100	.100				
X2	.100	.800	.100				
X3	.100	.100	.800				
Prev.	.333	.333	.333	.333	.333	.333	

First item has a logit of -0.5108256 from the group-level latent variable

References

Allison, K. R., Adlaf, E. M., Irving, H. M., Schoueri Mychasiw, N., & Rhem, J. (2016). The search for healthy schools: A multilevel latent class analysis of schools and their students. *Preventive Medicine Reports*, *4*, 331-337. doi: 10.1016/j.pmedr.2016.06.016

Asparouhov, T., & Muthén, B. O. (2015). Residual associations in latent class and latent transition analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 169-177. doi: 10.1080/10705511.2014.935844

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1-15. doi: 10.1348/000711002159617

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates. *Test*, *23*, 433-465. doi: 10.1007/s11749-014-0381-7

Bartolucci, F., Pennoni, F., & Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, Series A*, *170*, 115-132. doi: 10.1111/j.1467-985X.2006.00440.x

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, *41*, 164-171. doi: 10.1214/aoms/1177697196

Bennink, M., Croon, M. A., Keuning, J., & Vermunt, J. K. (2014). Measuring student ability, classifying schools, and detecting item bias at school level based on student-level dichotomous items. *Journal of Educational and Behavioral Statistics*, *39*, 180-201. doi: 10.3102/1076998614529158

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 397-460). Reading, MA: Addison-Wesley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodal inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261-304. doi: 10.1177/0049124104268644

Chen, K., & Kandel, D. B. (1995). The natural history of drug use from adolescence to the mid-thirties in a general population sample. *American Journal of Public Health*, *85*, 41-47. doi: 10.2105/AJPH.85.1.41

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, *79*, 762-771. doi: 10.2307/2288706

Clogg, C. C., & Goodman, L. A. (1985). Simultaneous latent structure analysis in several groups. *Sociological Methodology*, *15*, 81-110. doi: 10.2307/270847

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis with applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. (2012). Exploring dynamics in mood regulation: Mixture Latent Markov modeling of ambulatory assessment data. *Psychosomatic Medicine*, *74*, 366-376. doi: 10.1097/PSY.0b013e31825474cb

Dal Bianco, C., Paccagnella, O., & Varriale, R. (2016). A multilevel latent class analysis of the purchasing channels among European consumers. *Metron*, *74*, 293-309. doi: 10.1007/s40300-016-0100-0

Dias, J. G., Vermunt, J. K., & Ramos, S. (2015). Clustering financial time series: New insights from an extended hidden Markov model. *European Journal of Operational Research*, *243*, 852-864. doi: 10.1016/j.ejor.2014.12.041

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London, UK: Chapman & Hall.

Elliott, D. S., Huizinga, D., & Menard, S. (1989). *Multiple problem youth: Delinquency, substance use, and mental health problems*. New York, NY: Springer-Verlag.

Fagginger Auer, M. F., Hickendorff, M., Van Putten, C. M., Béguin, A. A., & Heiser, W. J. (2016). Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education*, *29*, 144-159. doi: 10.1080/08957347.2016.1138959

Ferdinand, R. F., De Nijs, P. F. A., Van Lier, P., & Verhulst, F. C. (2005). Latent class analysis of anxiety and depressive symptoms in referred adolescents. *Journal of Affective Disorders*, *88*, 299-306. doi: 10.1016/j.jad.2005.08.004

Fila, M. J., Paik, L. S., Griffeth, R. W., & Allen, D. (2014). Disaggregating job satisfaction: Effects of perceived demands, control, and support. *Journal of Business and Psychology*, *29*, 639-649. doi: 10.1007/s10869-014-9358-5

Finch, H. W., & Bronk, K. C. (2011). Conducting confirmatory latent class analysis using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 132-151. doi: 10.1080/10705511.2011.532732

Flory, K., Lynam, D., Milich, R., Leukefeld, C., & Clayton, R. (2004). Early adolescent through young adult alcohol and marijuana use trajectories: Early predictors, young adult outcomes, and predictive utility. *Development and Psychopathology*, *16*, 193-213. doi: 10.1017/S0954579404044475

Foli, K. J., South, S. C., Lim, E., & Jarnecke, A. M. (2016). Post-adoption depression: Parental classes of depressive symptoms across time. *Journal of Affective Disorders*, *200*, 293-302. doi: 10.1016/j.jad.2016.01.049

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102:6, December*, 460-465.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733-807.

Gilreath, T. D., Astor, R. A., Estrada, J. N., Johnson, R. M., Benbenishty, R., & Unger, J. B. (2014). Substance use among adolescents in California: A latent class analysis. *Substance Use & Misuse*, *49*, 116-123. doi: 10.3109/10826084.2013.824468

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psyhometrika*, *64*, 273-294. doi: 10.1007/BF02294296

Goodman, E., Maxwell, S., Malspeis, S., & Adler, N. (2015). Developmental trajectories of subjective social status. *Pediatrics*, *136*, e633-e640. doi: 10.1542/peds.2015-1300

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231. doi: 10.1093/biomet/61.2.215

Goodman, L. A. (2002). Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (p. 3-56). Cambridge, UK: Cambridge University Press.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p-values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337-350. doi: 10.1007/s10654-016-0149-3

Gunter, B., & Furnham, A. (1996). Biographical and climate predictors of job satisfaction and pride in organization. *Journal of Psychology: Interdisciplinary and Applied*, *130*, 193-208. doi: 10.1080/00223980.1996.9915001

Hagenaars, J., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Hamaker, E. L., Van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (p. 231-256). New York, NY: Routledge.

Harrell, P. T., Mancha, B. E., Petras, H., Trenz, R. C., & Latimer, W. W. (2012). Latent classes of heroin and cocaine users predict unique HIV/HCV risk factors. *Drug and Alcohol Dependence*, *122*, 220-227. doi: 10.1016/j.drugalcdep.2011.10.001

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (Second ed.). New York, NY: Routledge.

Isler, L., Liu, J. H., Sibley, C. G., & Fletcher, G. J. O. (2016). Self regulation and personality profiles: Empirical development, longitudinal stability and predictive ability. *European Journal of Personality*, *30*, 274-287. doi: 10.1002/per.2054

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research*, *25*, 137-155. doi: 10.1207/s15327906mbr2502_1

Kaplan, D., & Keller, B. (2011). A note on cluster effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 525-536. doi: 10.1080/10705511.2011.607071

Khalid, M. N., & Glas, C. A. W. (2016). Assessing item fit: A comparative study of frequentist and bayesian frameworks. *Measurement*, *90*, 549-559. doi: 10.1016/j.measurement.2016.05.020

Lambert, E. G., Hogan, N. L., & Barton, S. M. (2001). The impact of job satisfaction on turnover intent: A test of a structural measurement model using a national sample of workers. *Social Science Journal*, *38*, 233-250. doi: 10.1016/S0362-3319(01)00110-0

Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, *24*, 492-516. doi: 10.1177/0049124196024004004

Langeheine, R., & Van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research*, *18*, 416-441. doi: 10.1177/0049124190018004002

Laudy, O., Zoccolillo, M., Baillargeon, R. H., Boom, J., Tremblay, R. E., & Hoijtink, H. J. A. (2005). Applications of confirmatory latent class analysis in developmental psychology. *European Journal of Developmental Psychology*, *2*, 1-15. doi: 10.1080/17405620444000193

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & The interpretation and computation of some latent structures. In S. A. Stouffer (Ed.), *Studies in social psychology in World War II, Vol.4: Measurement and prediction* (p. 362-472). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science - Volume 3: Formulations of the person and the social context* (p. 476-543). New York, NY: McGraw-Hill Inc.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin Company.

Lee, T., Cai, L., & Kuhfeld, M. (2016). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 206-220. doi: 10.1080/10705511.2015.1014041

Liu, D., Mitchell, T. R., Lee, T. W., Holtom, B. C., & Hinkin, T. R. (2012). When employees are out of step with coworkers: How job satisfaction trajectory and dispersion influence individual- and unit-level voluntary turnover. *Academy of Management Journal*, *55*, 1360-1380. doi: 10.5465/amj.2010.0920

Lukočiené, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, *40*, 247-283. doi: 10.1111/j.1467-9531.2010.01231.x

Lukočiené, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (p. 241-250). Heidelberg, DE: Springer-Verlag.

Lundstedt, T., Seiferta, E., Abramob, L., Thelinc, B., Nyström, s., Pettersena, J., & Bergmana, R. (1998). Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems*, *42*, 3-40. doi: 10.1016/S0169-7439(98)00065-3

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490-504. doi: 10.1037/0033-2909.111.3.490

Mavridis, D., Moustaki, I., & Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (p. 135-162). Amsterdam, NL: Elsevier.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, 1-10. doi: 10.1186/1471-2288-7-34

Montgomery, D. C. (2012). *Design and analysis of experiments* (8th ed.). Hoboken, NJ: Wiley.

Muthèn, B. O., & Aspahourov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society, Series A*, *172*, 639-657. doi: 10.1111/j.1467-985X.2009.00589.x

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, *46*, 252-282. doi: 10.1177/0081175015581379

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2017). Power and type I error of local fit statistics in multilevel latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 216-229. doi: 10.1080/10705511.2016.1250639

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*, 45-60. doi: 10.1093/pan/mpt014

Oberski, D. L., Van Kollenburg, G. H., & Vermunt, J. K. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*, 267-279. doi: 10.1007/s11634-013-0146-2

Paas, L. P., Vermunt, J. K., & Bijmolt, T. H. A. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society Series A*, *170*, 955-974. doi: 10.1111/j.1467-985X.2007.00478.x

Park, J., & Yu, H.-T. (2016). The impact of ignoring the level of nesting structure in nonparametric multilevel latent class analysis. *Educational and Psychological Measurement*, *76*, 824-847. doi: 10.1177/0013164415618240

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5*, *L*, 157-175.

Pohle, J., Langrock, R., Van Beest, F. M., & Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models - Pitfalls, practical challenges and pragmatic solutions. *Journal of Agricultural Biological and Environmental Statistics*, *In press*, Obtained from https://arxiv.org/abs/1701.08673.

Poulsen, C. S. (1982). *Latent structure analysis with choice modeling applications*. Aarhus, DK: Aarhus University Press. (PhD Thesis)

R Development Core Team. (2015). *R v3.3: A language and environment for statistical computing*. Vienna, AT: R Foundation for Statistical Computing. (Computer Software)

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, DK: Danish Institute for Educational Research.

Roosma, F., Van Oorschot, W. J. H., & Gelissen, J. P. T. M. (2016). A just distribution of burdens? Attitudes toward the social distribution of taxes in 26 welfare states. *International Journal of Public Opinion Research*, *28*, 376-400. doi: 10.1093/ijpor/edv020

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 583-601. doi: 10.1080/10705510903203466

Samuel, R., Bergman, M. M., & Hupka-Brunner, S. (2013). The interplay between educational achievement, occupational success, and well-being. *Social Indicators Research*, *111*, 75-96. doi: 10.1007/s11205-011-9984-5

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., ... Miles, A. (2013). A new model of social class? Findings from the BBC's great British class survey experiment. *Sociology*, *47*, 219-250. doi: 0.1177/0038038513481128

Snijders, T. A. B., & Berkhof, J. (2008). Diagnostic checks for multilevel models. In J. de Leeuw & E. Mijer (Eds.), *Handbook of multilevel analysis* (p. 141-175). New York, NY: Springer.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Titman, A. C. (2007). *Model diagnostics in multi-state models of biological systems*. Cambridge, UK: University of Cambridge. (PhD Thesis)

Tomczyk, S., Hanewinkel, R., & Isensee, B. (2015). Multiple substance use patterns in adolescents: A multilevel latent class analysis. *Drug and Alcohol Dependence*, *155*, 208-214. doi: 10.1016/j.drugalcdep.2015.07.016

Van de Pol, F., & De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, *15*, 118-141. doi: 10.1177/0049124186015001009

Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, *20*, 213-247. doi: 10.2307/271087

Van Kollenburg, G. H. (2017). *Computer intensive methods for evaluating latent class model fit*. Tilburg, NL: Tilburg University. (PhD Thesis)

Van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2015). Assessing model fit in latent class analysis when asymptotics do not hold. *Methodology*, *11*, 65-79. doi: 10.1027/1614-2241/a000093

Van Mierlo, H. (2003). *Self-managing teamwork and psychological well-being*. Eindhoven, NL: Eindhoven University of Technology. (PhD Thesis)

Van Mierlo, H., Rutte, C. G., Kompier, M. A. J., & Doorewaard, H. A. C. M. (2005). Self-managing teamwork and psychological well-being: Review of a multilevel research domain. *Group and Organizational Management*, *30*, 211-235. doi: 10.1177/1059601103257989

Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, *47*, 247-275. doi: 10.1080/00273171.2012.658337

Vasdekis, V. G. S., Cagnone, S., & Moustaki, I. (2012). A composite likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika*, *77*, 425-441. doi: 10.1007/s11336-012-9264-6

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213-239. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2004). An em algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, *58*, 220-233. doi: 10.1046/j.0039-0402.2003.00257.x

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*, 33-51. doi: 10.1177/0962280207081238

Vermunt, J. K., Langeheine, R., & Böckenholt, U. (1999). Discrete time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*, 179-207. doi: 10.3102/10769986024002179

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2015). *Latent GOLD 5.0*. Belmont, MA: Statistical Innovations Inc. (Computer Software)

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (p. 373-385). Burlington, MA: Elsevier.

Von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research*, *2*, 29-48.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, *14*, 779-804. doi: 10.3758/BF03194105

Wiggins, L. M. (1955). *Mathematical models for the interpretation of attitude and behavior change: The analysis of multi-wave panel data*. Ann Arbor, MI: Columbia University. (PhD Thesis)

Wiggins, L. M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Amsterdam, NL: Elsevier.

Witteveen, D., & Attewell, P. (2017). The college completion puzzle: A hidden Markov model approach. *Research in Higher Education*, *58*, 449-467. doi: 10.1007/s11162-016-9430-2

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*, 53-82. doi: 10.1111/j.1467-9531.2007.00182.x

Zinn, A. (2015). A typology of supervision in child welfare: Multilevel latent class and confirmatory analyses of caseworker-supervisor relationship type. *Children and Youth Services Review*, *48*, 98-110. doi: 10.1016/j.childyouth.2014.12.004

Summary

This dissertation is aimed at introducing ways to locally test model assumptions of Multilevel Latent Class (LC) and Hidden or Latent Markov (LM) Models in order to improve their applicability. These models are relatively complex models that allow nested categorical data to be dealt with in a latent variable framework. As a result, they are frequently applied in the social sciences and data science fields where these classification methods have proven a valuable approach to analyze unobserved or unobservable phenomena. Yet, despite their value and applicability, fit testing for these models predominantly revolves around global fit statistics. Moreover, due to several technical aspects and their complex nature, true tests of global fit are often unobtainable, and model selection is primarily based on relative model fit statistics such as the Akaike and Bayesian Information Criterion.

Because both the multilevel LC and the LM model can be considered extensions to the regular LC model, the Bivariate Residual (BVR) that allows local assumption testing for the regular LC model formed the starting point of this project. By generalizing the idea of the BVR and make it applicable to multiple types of local dependence assumptions in the extended models, three goals are focused on. Namely providing relevant information about potential violations of the assumptions that these models make, make the statistics relatively easy to obtain, and inspecting the properties of these statistics.

In Chapter 2, two local fit statistics for the multilevel LC model are introduced. A regular LC model assumes that the indicator items that define the classification are conditionally independent when the latent variable is taken into account. This particular independence assumption is tested by the BVR that is available for the original model. By obtaining a Pearson residual on the observed and expected contingency tables containing the categories on two indicator items, the discrepancy between the observed and expected bivariate frequencies can be quantified. This discrepancy is indicative of violations of the conditional independence assumption, as perfect adherence to the assumption would mean that the expected and observed bivariate frequencies are identical and the residual would approach zero. Using this logic as the starting point two similar measures are constructed for the multilevel extension to the model. The assumptions that the extension adds are twofold,

namely that the model is able to adequately reproduce the within-group responses, and in order to deal with the nested data structure, that the individual observations are conditionally independent when taking into account the higher-level latent variable. Both of these issues are, or can be reduced to a form of assumed conditional independence. The BVR-group statistic deals with the first issue of the model adequately fitting the individual groups. Adhering to this assumption implies that the observed responses of individuals should be independent of their group membership, which is where it becomes apparent that this too is a conditional independence issue. The same approach as that of the BVR is applied, but rather than obtaining the residual between the observed and expected item-by-item contingency tables, the contingency tables concern the group-by-item association. Any indication of residual dependence means that observed group membership still affects the responses despite the assumption that the latent variable should capture all these effects. The BVR-pair statistic deals with the second issue of between respondent dependence, and similarly quantifies any discrepancy between the dependence that is present in the data, and the dependence that the model is able to reproduce. In order to accomplish this the pairwise response frequencies of all members in an observed group are obtained and compared to the model predicted pairwise frequencies, capturing any association among group-members that should be captured by the group-level latent variable. In an application it is shown that using these statistics leads to a better understanding of the origins of possible model misfit, and that by using the information provided by the statistics a globally better fitting model can be identified.

In Chapter 3 the properties of the in Chapter 2 introduced statistics are investigated using an extensive simulation study. Results indicate that the Type I error of the statistics is adequate, and power to indeed detect local assumption violations is generally high in situations where the data allows for multilevel LC modeling. Not only do the BVR-group and BVR-pair statistics detect the presence of misfit, they too are found to discriminate well between the problematic and non-problematic indicator items, as well as the different types of assumption violation that they are constructed to detect. Furthermore, the results indicate that the original BVR is limited to detecting misfit on the lower level of the model, and although BVR-group and BVR-pair can indicate higher-level misfit while its source is located on the lower level, using the three fit statistics in conjunction it is possible to precisely locate the misfitting areas in a large number of situations.

In Chapter 4, given the positive results in Chapters 2 and 3, similar local fit statistics are introduced for the LM model. These include a translation of the multilevel LC statistics to deal with longitudinally nested data, as well as the introduction of

additional measures to capture autoregressive dependence. The BVR-case and BVR-time statistics introduced in this chapter are broadly comparable to the BVR-group statistic. However, where the assumption in the multilevel model entails reproducing the univariate item distributions within each of the observed groups, the required reproduction for a LM model can be considered in two directions. That is, the response pattern of one case over time, and the individual responses within each measurement occasion need to be adequately reproduced by the model. By using the case-by-item and time-by-item contingency tables, instead of the group-by-item frequencies, residuals for both these issues can be obtained and are indicative of any discrepancies between the model based frequencies and the observed frequencies. The BVR-pair statistic for the LM model is comparable to the BVR-pair statistic for multilevel LC models in a similar fashion. Here too the important aspect is the assumed independence between multiple observations of a single case, which can be inspected with an almost analogous approach to testing the assumed independence of multiple observations within an observed group. Because the BVR-pair statistic in this instance tests all possible residual dependence between all possible combinations of measurement occasions it can be split up into different BVR-lag statistics, testing the conditional independence assumption violations between measurement occasions with a fixed distance between them. This provides information that, when combined with the BVR-pair measure for LM models can indicate whether the Markov assumption holds, and if not whether this is due to the Markov assumption being violated for longer or shorter periods of lag. Applying these statistics to two data sets it is shown that they are indeed able to provide a lot of additional information on model misfit, and allow improvements to the model to be made. Moreover, by using all the presented BVR residuals in conjunction enough information is available to indicate whether the model is particularly well-suited to modeling the between-person or between-time differences.

In Chapter 5 it is attempted to make testing for local misfit, that is obtaining a p-value for the residuals, less intrusive. Because obtaining the bootstrap p-values requires the re-estimation of the model on several hundred replicate data sets, the time it takes to compute the final results of the analysis increases rapidly with model complexity. By avoiding the step of model estimation during the bootstrap procedure a large part of the computational complexity can be avoided. The way in which this is achieved is by only considering whether a particular aspect of the data is correctly reproduced. Although answering a slightly different question than the BVR residuals, the conditional dependence that the BVR statistics test for can be reformulated into statistics that are directly obtainable from the data. By subsequently

generating replicate data from the model as would be done for a regular bootstrap, but instead building the distribution of the statistic of interest directly from the data without model estimation it can be inspected whether the model could be the data generating model for that particular aspect of the data. Although further study is required, the initial results of this approach are positive and generally similar types and severity of misfit can be detected when compared to the BVR statistics.

Acknowledgments

Over the years there are many people that have contributed to this thesis, either directly or indirectly by being supportive, friendly, and amazing people. The list is simply too long to merely even mention everyone, let alone thank you all in the detail that I'd like. I'll have a stab at it, and have you know that most of these mentions are way too brief to convey how much I value having you around.

Academically, it is self-evident that my promotor and co-promotor have been extremely important. I want to thank both of you wholeheartedly for four years of encouragement, supervision and all the help you provided. However, it did not end with mere academic supervision, and your personal support has been invaluable. Jeroen, thank you for all the insights, space, and nudges in the right direction you have given me. Over the years my appreciation for you as a person has greatly increased in addition to the appreciation of you as a supervisor. Daniel, your door was always open and especially in the first years of this project you always took the time for extensive explanations, introducing the notion that most of the 'bears I saw on the road' were actually sitting on the hard shoulder, and the occasional little boost to my self-confidence.

Further help was always generously provided by the members of our lab group. I have learned a lot from the papers I have read and the feedback I received from the 'Extended Vici Group.' Thanks for all the effort Margot, Zsuzsa, Daniel P., Dereje, Davide, Fetene, Mattis, Geert, Lianne, Katrijn, Reza, Laura, Kim, Niek, Leonie, and Pia.

Of course, I owe a lot to all colleagues, many of whom became good friends and (drinking) buddies. In the first months I was impressed by the amount of gossip Margot and Zsuzsa managed to accumulate and willing to share and the friendly and calm advice of Pieter. I felt welcomed by the group of PhDs that started around the same time and I spent the past four years enjoying vast amounts of jokes, coffee, frustrations, tips and tricks, drinks, adventures, banter, and/or good conversations with. Michèle, Robert, Paulette, Coosje, Florian, Robbie, Ingrid, Chris, Davide, and Eva, thanks for the amazing time. Over the years this list of awesome and friendly people quickly grew longer with Eva, Sara, Niek, Jules, Laura, Zhengguo, Hilde, Leonie, Dino, JJ, Elise, Paul, Esther, Shuai, Andrea and Hannah joining the ranks.

A special mention goes out to my conference crew, officemate, and paranymphs. Geert 'the bald' and Mattis 'the third', thanks for all the laughs, adventures, late night talks, and being witness to some of the weirdest conference moments. Lianne, thanks for all the support and friendship over many years, from the afterparties at

Aaron to the honor of being your paranymph. Also, to settle the matter once and for all I will put it in writing: The sole cause of me falling of my bike was you yanking the jacket out of my hands. Luc, I have always greatly appreciated you as a lecturer during my time as a student and am happy that I can now also say that I greatly appreciate you as a buddy and colleague.

If I had to list all others that I did not mention this would simply become a list of all current and former members of our department, supplemented by a list of PhD students from other fields, as I have enjoyed the company of any and all of you. I firmly believe it would be nigh on impossible to find a group of people that I would enjoy seeing every day and make me feel at home in the office this much. Thanks, and I am incredibly happy to get to see more of you in the coming years!

Thanks also go out to my family. Dad, Adrie, I have already once attempted to explain what you and mom mean to me. Every choice that you two made when I was younger, and any decision we have made together has led to loving, joyful and proud memories. Sister, Cynthia, I am going to need very few words for this: "Tsja, een kleinigheidje blijf je houden, maar dat kunnen wij samen eens eventjes." Mom, Irma, without you I'd never have been able to achieve any of this. Without you I would not have been able to achieve a lot of things. Without you I think the three of us would have been quite lost in the world. Thanks for all the life lessons, big and small. The bottomless optimism, endless willpower and heartfelt kindness with which you led life is inspiring. Ik mis je.

Of course, there are also friends to thank outside of work, whose importance in my life and their help in pulling me through some hard times cannot be overstated. Martijn 'Wuis' and Annemarie 'Almo', thank you for the honor of arranging the most expensive day of your lives, for all the pineapples, and all the joy. Emma and Paul, thank you for the open house policy, the long conversations, and all the social awkwardness that makes for good stories.

And then, then there's this improbable group of friends, that against all social conventions still sticks together. Bob 'Bee' and Germa 'Gee', thanks for the whiskey by the fire and the conversations that follow effortlessly. Kelly 'Driehoek' and Gino 'Shibz', thanks for the entertainment, the razor-sharp sarcasm where needed, and the hugs where needed. Suzanne 'Suus' and Michail 'Finch', thanks for a second home. Regardless of whether I could bike there or it was a three-hour drive, it always felt like leaving home to arrive at home. CJ 'Pluis', no need to get all mushy after eighteen years, just make sure there is enough beer. Yvette 'Ief' and Robbin 'Vettel' thanks for the support and hugs in times they are needed the most. Jark and Suzanne 'Suus', thanks for all the kind words and long talks on mutual dislikes. Pieter 'Mimic' and Susanne 'Suus', thanks for the relaxing evenings and your down to earth vision on pretty much everything. Marjet 'Jetje', even though our roads are no longer one and the same, it'd be weird to ignore you here, because there is so much support and love to thank you for after sharing fourteen years of happiness and sorrow. I hope your memories of our time together are as warm as mine.

Mamma,

In je hoofd,

kun je alles.

Fietsen naar de maan,

op de wolken staan.

Strelen met je handen los,

lopen door een donker bos.

Vechten als een tijger,

dansen met een elf.

Afscheid nemen zonder tranen.

Alles gaat vanzelf.

~Theo Olthuis

Annemarie,

De macht van het kleine gaf jou juist de kans,

De kracht van het kleine tentoon te spreiden.

$$\chi + \square = \chi^2$$

Or as the unofficial undertitle would have it: *Millions of chi-square values*